Applying Factor Analysis and Item Response Models to Undergraduate Statistics Exams

A Master Thesis Presented

by

Vinh Hanh Lieu

Tester: **Prof. Dr. Wolfgang Härdle**

Director: **Dr. Sigbert Klinke**

Humboldt University, Berlin

Economics Faculty

Institute of Statistics and Econometrics

Berlin, February 13, 2013

Declaration of Authorship

I hereby confirm that I have authored this master thesis independently and without use of others than the indicated resources. All passages, which are literally or in general matter taken out of publications or other resources, are marked as such.

Vinh Hanh Lieu

Berlin, February 13, 2013

Abstract

Item response theory is a modern model-based measurement theory. More recently, the popularity of the IRT models due to many important research applications has become apparent. The purpose of conducting explanatory and confirmatory factor analysis is to explore the interrelationships between the observed item responses and to test whether the data fit a hypothesized measurement model. The item response models applied to undergraduate statistics exams show how the trait level estimates from the models depend on both examinees´ responses and on the properties of the administered items. The reliability analysis indicates that both exams measure a single unidimensional latent construct, ability of examinees very well. Two-factor model is obtained from the explanatory factor analysis of the second exam. Based on several goodness of fit indices confirmatory factor analysis verifies again the obtained results in the explanatory factor analysis. We fit the testlet-based data with the dichotomous and polytomous item response models. Estimated item parameters and total information functions from the three different models are compared with each other. Difficulty item parameters estimated from one and two parameter logistic item response functions correlate highly. The first statistics exam is a good test measurement since examinees with all level of abilities measured with the questions of different difficulty along the whole scale. Several more difficult questions are needed to measure high-proficiency examinees in the second round. The polytomous IRT model provides more information than the two parameter logistic item response model only in high-ability level.

Keywords:
exploratory factor analysis, confirmatory factor analysis, dichotomous and polytomous item response theory, IRTPRO 2.1

Acknowledgments

First, I would like to acknowledge my supervisor, Dr. Sigbert Klinke, who has supported me a lot during the writing of this work.

Second, my gratitude to Prof.Dr. Wolfgang Härdle for instructing me to the world of economic statistics is sincere.

A big thank goes to H. Wainer, E. Bradlow and X. Wang, who provided computer program Scoright 3.0, an essential tool for testlet models.

I am also very grateful to my father, my mother and my husband for their warmest encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Test or exam is an instrument used when one wants to measure something. Lecturers always have to answer many questions before giving a test. What information about examinees testers want to know? Which tasks should be given to examinees? How can the test be scored? How well does the test score work if the assumptions of the structure of the test are violated? The demands on the measurements of test scores have increased. Until now, there have been many ways of scoring tests based on characteristics of tests.

In this thesis exploratory and confirmatory factor analysis, dichotomous and polytomous item response theory (IRT) will be proposed. IRT describes what happens when an item meets an examinee. The assumption of dichotomous model is conditionally local independence within items. Unfortunately, it does not hold for several tests, for example a reading passage and a set of associated items (testlet). In order to make the local dependencies within testlets disappear, we can consider the entire testlet as a unit and score it polytomously (Wainer, 2007). That is the purpose of employing polytomous IRT.

The collected data is from the two statistics exams for undergraduate students of School of Business and Economics bzw. Ladislaus von Bortkiewicz Chair of Statistics, Humboldt University, Berlin in summer semester 2011. The examinees are 176 and 171 in the first exam (26.07.2011) and the second exam (13.10.2011).

The softwares used in this work are IRTPRO for Windows 2.1, Scoright 3.0, R and M-plus. IRTPRO for Windows 2.1 is developed by Li Cai, David Thissen & Stephen du Toit. This product has replaced the four programs Bilog-MG, Multilog, Parscale and Testfact which overlapp somehow in functionalities. The program can be used with Windows7, Vista and XP operating systems.

The overview of data will be given in the next chapter. Chapter 3 introduces several applied statistical methods, such as reliability analysis, exploratory and confirmatory factor analysis, the methods of dichotomous and polytomous item response theory. Chapter 4 gives us an introduction about the software IRTPRO 2.1. In subsequent chapters the empirical results of the above-mentioned methods will be presented. Several conclusions will be drawn in the last chapter.

Chapter 2

Overview of Data

The data used in this thesis is extracted from the two statistics exams for undergraduate students of School of Business and Economics bzw. Ladislaus von Bortkiewicz Chair of Statistics, Humboldt University, Berlin in summer semester 2011.

The data consists of 176 examinees in the first exam, 171 examinees in the second exam. The students major mainly in Betriebswirtschaftslehre (BWL) and Volkswirtschaftslehre (VWL). The data were divided into three groups. These are BWL_BA (BA for bachelor), VWL_BA and Other. BWL_BA students are about 50% of data in both exams. 27,3% and 33,9% of data are VWL_BA students in the first and second exams. The other groups are 22,1% and 17%, respectively. 77,3% and 75,4 % examinees have passed in the first and second round.

Both exams have six exercises. Table 2.1 and 2.2 showed the characteristics of two exams. The structure of the two exams is really similar. They are combinatorics, probability, univariate variable, bivariate variables and distribution function in theoretical as well as practical field. There are a total of 28 questions in the first exam and 23 questions in the second one.

The first question of exercise 4 in the first exam seems to be the easiest question with 87% correct answers. It is a question about bivariate variables. The exercise 6 in the first and second round is the most difficult exercise with several questions having very low percent of right answers. In the second round, the first two questions of exercise 3 are the easiest ones with 94% correct answers. They are about univariate variable. The answers of several questions in each exercise are dependent on the previous answers. Most models are often applied with the assumption of the independence within items. How can we handle with this problem? The solution will be introduced in the following chapters.

The data have been dichotomized with the values 0 and 1. Each answer got equal and more than fifty percent of maximal points achieves a value 1, otherwise 0. The answer of the question with the codification of 1 is considered as correct and vice versa.

No.	Question	Theory/Practice	Field	Percent of correct answer
1	Q11_R1	Theory	Combinatorics	0,71
2	Q12_R1	Theory	Combinatorics	0,71
3	Q13_R1	Theory	Combinatorics	0,46
4	Q21_R1	Theory	Probability	0,71
5	Q22_R1	Theory	Probability	0,74
6	Q23_R1	Theory	Probability	0,62
7	Q24_R1	Theory	Probability	0,59
8	Q25_R1	Theory	Probability	0,46
9	Q26_R1	Theory	Probability	**0,22**
10	Q31_R1	Practice	Univariate	0,78
11	Q32_R1	Practice	Univariate	0,81
12	Q33_R1	Practice	Univariate	0,75
13	Q34_R1	Practice	Univariate	0,37
14	Q35_R1	Practice	Univariate	0,51
15	Q36_R1	Practice	Univariate	0,68
16	Q37_R1	Practice	Univariate	0,57
17	Q41_R1	Practice	Bivariate	**0,87**
18	Q42_R1	Practice	Bivariate	0,51
19	Q43_R1	Practice	Bivariate	0,59
20	Q44_R1	Practice	Bivariate	0,81
21	Q51_R1	Practice	Bivariate	0,52
22	Q52_R1	Practice	Bivariate	0,68
23	Q61_R1	Theory	Univariate	0,61
24	Q62_R1	Theory	Univariate	0,43
25	Q63_R1	Theory	Univariate	0,36
26	Q64_R1	Theory	Univariate	**0,27**
27	Q65_R1	Theory	Univariate	**0,19**
28	Q66_R1	Theory	Distribution	**0,22**

Table 2.1: Characteristics of the first exam and the percent of correct answer of each question

No.	Question	Theory/Practice	Field	Percent of correct answer
1	Q11_R2	Theory	Combinatorics	0,81
2	Q12_R2	Theory	Probability	0,90
3	Q21_R2	Theory	Probability	0,68
4	Q22_R2	Theory	Probability	0,74
5	Q31_R2	Practice	Univariate	**0,94**
6	Q32_R2	Practice	Univariate	**0,94**
7	Q33_R2	Practice	Univariate	0,57
8	Q34_R2	Practice	Univariate	0,49
9	Q35_R2	Practice	Univariate	0,45
10	Q36_R2	Practice	Univariate	0,72
11	Q37_R2	Practice	Univariate	0,75
12	Q41_R2	Practice	Bivariate	0,90
13	Q42_R2	Practice	Bivariate	0,63
14	Q43_R2	Practice	Bivariate	0,61
15	Q44_R2	Practice	Bivariate	0,43
16	Q51_R2	Theory	Bivariate	0,54
17	Q52_R2	Theory	Bivariate	0,64
18	Q61_R2	Theory	Univariate	0,75
19	Q62_R2	Theory	Univariate	0,65
20	Q63_R2	Theory	Univariate	0,52
21	Q64_R2	Theory	Univariate	0,42
22	Q65_R2	Theory	Univariate	0,49
23	Q66_R2	Theory	Distribution	**0,11**

Table 2.2: Characteristics of the second exam and the percent of correct answer of each question

Chapter 3

Applied statistical methods

3.1 Reliability analysis

Reliability is the correlation between the observed variable and the true score when the variable is an inexact or imprecise indicator of the true score (Cohen and Cohen, 1983). Inexact measures may come from guessing, differential perception, recording errors, etc. on the part of the observers.

Cronbach´s alpha is a coefficient of reliability or internal consistency of a latent construct. It measures how well a set of items or variables measures a single unidimensional latent construct. When data have a multidimensional structure, Cronbach's alpha will usually be low. Cronbach´s alpha is calculated with the formula

$$\alpha = \left(\frac{m}{m-1}\right)\left(1 - \frac{\sum_{j=1}^{m} S_j^2}{S_Y^2}\right) \tag{3.1}$$

where

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$$

$$\overline{Y} = \sum_{j=1}^{m}\overline{X}_j$$

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} X_{ij} - \overline{Y}\right)^2$$

- n is the number of observations
- m is the number of items in scale
- S_j^2 is the variance of item j
- S_Y^2 is the variance of total score

The standardized Cronbach´s alpha can be written as a function of the number of items m and the average inter-correlation $\overline{r}$ among the items.

$$\alpha = \frac{m\overline{r}}{1 + (m-1)\overline{r}} \tag{3.2}$$

From this formula one can see that if one increases the number of items, Cronbach's alpha will rise. Additionally, if the average inter-item correlation is low, alpha will be low and reversed. The range of the alpha is from 0 to 1. A reliability coefficient of 0,70 or higher is considered "acceptable" in most research situations.

3.2 Exploratory factor analysis

Exploratory factor analysis (EFA) is used to determine continuous latent variables which can explain the correlations among a set of observed variables. The continuous latent variables are referred to as factors. The observed variables are referred to as factor indicators. In our data, factor indicators are dichotomized variables. The basic objectives of EFA are:

- Explore the interrelationships between the observed variables
- Determine whether the interrelationships can be explained by a small number of latent variables

The introduction and definition of EFA in this section were extracted from the book of Härdle and Simar (2003). A p-dimensional random vector X with mean μ and covariance matrix Σ can be represented by a matrix of factor loadings Q ($p \times k$) and factors F ($k \times 1$). The number of factors, k should always be much smaller than p.

$$X_{(p\times 1)} = Q_{(p\times k)}.F_{(k\times 1)} + \mu_{(p\times 1)} \tag{3.3}$$

If $\lambda_{k+1} = ... = \lambda_p = 0$ (eigenvalues of Σ), one can express X by the factor model 3.3. At that time, Σ will be a singular matrix. In practice this is rarely the case. Thus, the influences of the factors are often split into common (F) and specific (U) ones. U is a ($p \times 1$) matrix of the random specific factors which capture the individual variance of each component. The random vectors F and U are unobservable and uncorrelated.

$$X_{(p\times 1)} = Q_{(p\times k)}.F_{(k\times 1)} + U_{(p\times 1)} + \mu_{(p\times 1)} \tag{3.4}$$

It is assumed that:

- $EF = 0$,
- $Var(F) = I_k$,
- $EU = 0$,
- $Cov(U_i, U_j) = 0$, $i \neq j$
- $Cov(F, U) = 0$

EFA is normally performed in four steps (Klinke and Wagner, 2008):

1. estimating the correlation matrix $\hat{R}$ between the p variables. One uses Bravais-Pearson correlation coefficient, if the data is metrical. For ordinal data Kendall´s τ_b, Spearmans rank correlation or polychoric correlation can be applied.

2. estimating the number of common factors k, k < p

3. estimating the loading matrix $\hat{Q}$ of the common factors. There are different methods for computation of factor model: Principal Component (PC), Principal Axis (PA), Maximum Likelihood (ML) and Unweighted Least Squares (ULS).

4. rotation of the factor loadings helps to interpret the factors easier, e.g. Varimax, Promax.

3.2.1 Tetrachoric correlation

In our case, tetrachoric correlation matrix for binary variables will be calculated. Tetrachoric correlation is used when both variables are dichotomies which are assumed to represent underlying bivariate normal distributions. Tetrachoric correlation can be a nonpositive definite correlation matrix when one of eigenvalue is negative. It may reflect the violation of normality, outliers, or multi collinearity of variables.

Let y_i, i=1, ..., n be a binary response, y_i^* be a corresponding continuous latent response variable. The formulation is closely related to the ordinary factor analysis model for quantitative variables.

$$y_{ij} = \begin{cases} 1, & \text{if } y_{ij}^* \geq \tau_j \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

$$y_i^* = \nu + \Lambda\eta_i + \epsilon_i \tag{3.6}$$

where

- $j = 1, ..., p$ refers to the observed dependent variable
- τ is threshold parameter
- ν is a p-dimensional parameter vector of measurement intercepts
- Λ is a $p \times m$ parameter matrix of measurement slopes or factor loadings
- η is an m-dimensional vector of latent variables (contructs or factors)
- ϵ is a p-dimensional vector of residuals or measurement errors

The structural part of the model is given by

$$\eta_i = \alpha + B\eta_i + \Gamma x_i + \zeta_i \tag{3.7}$$

where

- $\alpha_{m\times 1}$ is vector of latent intercepts
- $B_{m\times m}$ is a matrix of dependent latent variable slopes with zero diagonal elements, assumed that I-B nonsingular
- $\Gamma_{m\times q}$ is a matrix of covariate slopes
- x_i is a vector of observed covariates

- ζ_i is a vector of latent variable residuals

The mean vector μ_i^* and covariance matrix Σ_i^* of y_i^* are derived under three assumptions:

- ϵ_i are i.i.d distributed with mean zero and diagonal covariance matrix Θ
- ζ_i are i.i.d distributed with mean zero and covariance matrix Ψ
- ϵ_i and ζ_i are uncorrelated

$$\mu_i^* = \Lambda(I-B)^{-1}\alpha + \Lambda(I-B)^{-1}\Gamma x_i \tag{3.8}$$

$$\Sigma_i^* = \Lambda(I-B)^{-1}\Psi(I-B)^{-1}\Lambda^T + \Theta \tag{3.9}$$

Let μ_{ij} denote mean of y_{ij} given x_i

$$\begin{aligned} \mu_{ij} &= E(y_{ij}|x_i) = 1 \cdot P(y_{ij}=1|x_i) + 0 \cdot P(y_{ij}=0|x_i) \\ &= P(y_{ij}^* > \tau_j|x_i) \\ &= \int_{\tau_j}^{\infty} f(y;\mu_{ij},\sigma_{ijj}^*)dy \end{aligned} \tag{3.10}$$

Because the variance of y_{ij}^* is not identifiable when binary data is observed. It is assumed that Σ_i^* has unit diagonal elements, hence $\sigma_{ijj}^* = 1$, j=1, ..., p. It follows that

$$\begin{aligned} \mu_{ij} &= \int_{\tau_j-\mu_{ij}^*}^{\infty} \phi(z)dz \\ &= \Phi(-\tau_j + \mu_{ij}^*) \end{aligned} \tag{3.11}$$

and the conditional correlation of y_{ij} and y_{ik} given by σ_{ijk}

$$\sigma_{ijk} = E(y_{ij}y_{ik}|x_i) - \mu_{ij}\mu_{ik} \tag{3.12}$$

where

$$\begin{aligned} E(y_{ij}y_{ik}|x_i) &= 1 \cdot P(y_{ij}=1, y_{ik}=1|x_i) + 0 \\ &= P(y_{ij}^* > \tau_j, y_{ik}^* > \tau_k|x_i) \\ &= \int_{\tau_j-\mu_{ij}^*}^{\infty}\int_{\tau_k-\mu_{ik}^*}^{\infty} g(z_1,z_2|x_i;\sigma_{ijk}^*)dz_1dz_2 \\ &= \Phi^*(-\tau_j+\mu_{ij}^*; -\tau_k+\mu_{ik}^*; \sigma_{ijk}^*) \end{aligned} \tag{3.13}$$

3.2.2 Estimation of common factors

The aim of factor analysis is to explain variations and covariations of multivariate data using fewer variables, the so-called factors. The unobserved factors are much more interesting than the observed variables themselves. How do the estimation procedures look like?

The estimation of factor model is based on the covariance or correlation matrix of data. Using the assumptions of factor model, the covariance matrix of data X can be shown as follows. After standardizing the observed variables, the correlation of X is computed with the similar form as the covariance matrix of X.

$$\begin{aligned}\Sigma &= E(X-\mu)(X-\mu)^T = E(QF+U)(QF+U)^T \\ &= QE(FF^T)Q^T + E(UU^T) \\ &= QVar(F)Q^T + Var(U) \\ &= QQ^T + \Psi \end{aligned} \tag{3.14}$$

The objective of FA is to find the loadings Q and the specific variance Ψ which are deduced from the covariance 3.14. The factor loadings are not unique. Taking the advantage of this non-uniqueness, we get a new loadings matrix (multiplication by an orthogonal matrix) which can make interpretation of factors easier as well as more understandable.

Factor loadings matrix Q gives the covariance between observed variables X and factors F. It is very meaningful to find another rotation which can show the maximal correlation between factors and original variables.

3.2.3 Principal component analysis (PCA)

The objective of PCA is to reduce the dimension of multivariate data matrix X achieved through linear combinations. The first principal component (PC) chosen captures the highest variance of the data whose direction is the eigenvector γ_1 corresponding to the largest eigenvalue λ_1 of the covariance matrix Σ. Orthogonal to the direction γ_1 we find the second PC with the second highest variance.

We centered the variable X to obtain a zero mean PC variable Y

$$Y = \Gamma^T(X-\mu) \tag{3.15}$$

The variance of Y will be equal to the eigenvalue Λ. The components of the eigenvectors are the weights of the original variables in the PCs.

The principal component method in factor analysis can be done as follows:

- spectral decomposition of empirical covariance matrix $S = \Gamma\Lambda\Gamma^T$
- approximation loadings $\hat{Q} = [\sqrt{\lambda_1\gamma_1}, ..., \sqrt{\lambda_k\gamma_k}]$ where k is number of factors
- estimation of specific variances by $\hat{\Psi} = S - \hat{Q}\hat{Q}^T$

Residual matrix analytically achieved from principal component solution so that it is smaller than the sum of the failing eigenvalues.

$$\sum_{i,j}(S - \hat{Q}\hat{Q}^T - \hat{\Psi})^2_{ij} \leq \lambda^2_{k+1} + ... + \lambda^2_p$$

3.2.4 Number of extracted factors

Kaiser criterion

According to the Kaiser criterion, factors should be extracted when their eigenvalues are bigger than one. An eigenvalue of a factor indicates variance of all variables which are explained by the factor.

Horn´s Parallel analysis

Horn´s Parallel analysis compares the eigenvalues obtained from empirical correlation matrix with those obtained from normal distributed random variables. The number of extracted factors corresponds to the number of non-random eigenvalues that are above the distribution of eigenvalues derived from random data (Bortz, 1999, 229). See Bortz, 1999 for more details.

3.2.5 Rotation of factors

Varimax rotation method proposed by Kaiser (1985) is an orthogonal rotation of the factor axes which maximizes the variance of the squared loadings of a factor on all the variables in a factor matrix. Promax rotation rotates the factor axes allowing to have an oblique angle between them. A rotated solution helps to identify each variable with a single factor and makes the interpretation of the factors easier.

3.3 Confirmatory factor analysis

Confirmatory factor analysis (CFA) is used to test whether the data fit a hypothesized measurement model proposed by a researcher. This hypothesized model is based on theory or previous study. The difference to EFA is that each variable is just loaded on one factor. Error terms contain the remaining influence of variables. The null hypothesis is that the covariance matrix of the observed variables is equal to the estimated covariance matrix.

$$H_0 : S = \Sigma(\hat{\theta})$$

where $\Sigma(\hat{\theta})$ is the estimated covariance matrix.

3.3.1 Estimation of model parameter

Muthén (1984) considered the weighted least square (WLS) fitting function as follows. An advantage of WLS-discrepancy function is that the assumption about skewness and kurtosis is not needed. Since these information are considered in the so-called asymptotic variance-covariance matrix W.

$$F_{WLS} = (s - \sigma(\theta))^T W^{-1} (s - \sigma(\theta)) \tag{3.16}$$

where

- s is a vector of elements in empirical covariance matrix
- $\sigma(\theta)$ is a vector of corresponding elements in estimated covariance matrix

- W is a covariance matrix of variance and covariance of measured variables.

σ is obtained by multivariate regression of p-dimensional vector y on q-dimensional vector x. The estimation of unknown parameters in this regression is carried out in two steps. Consider as an example the case of two binary variables y_1 and y_2 regressed on x.

Univariate-response probit regression (UPR) 3.11 with log likelihood l_{ij} for individual i and variable j is computed as follows

$$l_{ij} = y_{ij} \log P(y_{ij} = 1|x_i) + (1 - y_{ij}) \log P(y_{ij} = 0|x_i) \tag{3.17}$$

Bivariate-response probit regression (BPR) 3.13 with log likelihood l_{ijk} for individual i and variables j, k is

$$\begin{aligned} l_{ijk} = y_{ij}y_{ik} \log P(y_{ij} = 1, y_{ik} = 1|x_i) + \\ y_{ij}(1 - y_{ik}) \log P(y_{ij} = 1, y_{ik} = 0|x_i) + \\ (1 - y_{ij})y_{ik} \log P(y_{ij} = 0, y_{ik} = 1|x_i) + \\ (1 - y_{ij})(1 - y_{ik}) \log P(y_{ij} = 0, y_{ik} = 0|x_i) \end{aligned} \tag{3.18}$$

Solve the following equation to achieve parameters

$$\sum_{i=1}^{n} \partial l(i)/\partial\sigma = 0 \tag{3.19}$$

- From maximum-likelihood estimates we will receive threshold parameter τ, coefficients of y (probit slopes) in UPR.
- In the second step, ρ (residual covariance for y_j and y_k) are obtained in BPR, holding τ and probit slopes fixed at the estimated values from the UPR.

3.3.2 Tests of model fit

The χ^2 test statistic, goodness of fit indices such as RMSEA, TLI and CFI are used to evaluate to what extent a particular factor model explains the empirical data (Muthén, 2004).

Chi-square test

The χ^2 test checks the hypothesis that the theoretical covariance matrix corresponds to the empirical covariance matrix. The test statistic is χ^2-distributed under the assumption of the null hypothesis. The null hypothesis will be rejected if the value of the test statistic is large. It is not used for the large samples because big n will make the test always significant.

$$\chi^2 = (n - 1)F(S, \Sigma(\hat{\theta})) \tag{3.20}$$

where

- n is the number of observations
- F is the minimum of the discrepancy.

Root Mean Square Error of Approximation (RMSEA)

RMSEA is a measure for the model deviation per degree of freedom which ranges from 0 to 1. A value less than 0,05 indicates a good model fit. The values which are higher than 0,50 are indicative of bad model fit. The model is unacceptable when the value is bigger than 0,1.

$$RMSEA = \sqrt{\frac{\chi^2/df - 1}{n - 1}} \tag{3.21}$$

Comparative Fit Index (CFI)

CFI measures the discrepancy between the data and the hypothesized model. In null model the correlation and covariance of the latent variables are assumed equal to 0. A CFI value of 0,95 or larger indicates a good model fit.

$$CFI = \frac{(\chi_0^2 - df_0) - (\chi_1^2 - df_1)}{\chi_0^2 - df_0} \tag{3.22}$$

where

- χ_0^2 is chi-square value of null model (all parameters are set to zero)
- df_0 is degrees of freedom of null model
- χ_1^2 is chi-square value of the hypothesized model
- df_1 is degrees of freedom of the hypothesized model

Tucker Lewis Index (TLI)

TLI also called Non Normed Fit Index has the same meaning as CFI. TLI should range between 0 and 1, with a value of 0,95 or greater indicating a good model fit. The index is not influenced by the size sample. So one can use it for the data with many observations.

$$TLI = \frac{\chi_0^2/df_0 - \chi_1^2/df_1}{\chi_0^2/df_0 - 1} \tag{3.23}$$

3.4 Dichotomous Item Response Theory

3.4.1 Introduction

Item response theory (IRT) is a latent trait theory. The theory models the response of an examinee of given ability to each item of the test. IRT provides the probability of a correct answer to an item. It is a mathematical function of person and item parameters. The person parameter is also called latent trait or proficiency. Examinees at higher levels of θ have a higher probability of responding correctly an item. Item parameters may contain difficulty (b), discrimination (a), and pseudoguessing (c) parameters. The definition and explanation of various IRT models extracted from Ayala (2009), Wainer and Bradlow (2007) are presented in this section.

IRT has a number of advantages over clasical test theory methods. First, IRT item parameters are not dependent on the sample size. Second, IRT models measure scale precision across the underlying latent variable. Third, the person´s trait level is independent of the questions in the scale. Three assumptions are needed for this model:

1. A unidimensional trait θ;
2. Local independence of items;
3. The relationship between the latent trait and item responses presented by monotonic logistic function.

The unobservable construct or trait is measured by the questionnaire, e.g. anxiety, physical functioning, ability of examinees, ... The trait is assumed to be measurable on a scale with a mean of 0.0 and a standard deviation of 1.0. With a given proficiency local independence of items means that the probability of answering an item correctly is independent of responses to any of the other items.

3.4.2 Model Specification

Item response function (IRF) gives the probability that a person answers an item correctly with a given proficiency. Persons with high ability have more chance to answer items correctly than persons with lower ability. The probability depends on item parameters of the IRF. Or we can say that the item parameters determine the shape of the IRF. In this section, we will present the three basic models of IRT.

One parameter logistic item response function

$$P_{ij}(\theta_i, b_j) = \frac{exp(\theta_i - b_j)}{1 + exp(\theta_i - b_j)} \tag{3.24}$$

where the index $i = 1, ..., n$ refers to the person, the index $j = 1, ..., m$ refers to the item and $P_{ij}(\theta_i, b_j)$ is the probability with proficiency θ_i responding correctly to an item of difficulty b_j. Figure 3.24 is a plot of 1PL model also called Rasch model. The person´s ability is denoted with θ and it is plotted on the horizontal axis. Three items of different difficulty ($b = -1, 0, 1$) are illustrated. The item characteristic curves (ICCs) or trace lines for this model are parallel to one another. When we increase the difficulty parameter b, the item response function will move to the right. The values of this function are between 0 and 1 for any argument between $-\infty$ and $+\infty$. This makes it appropriate for predicting probabilities.

The 1PL model predicts the probability of a correct response from the interaction between the individual ability θ and the item parameter b. The horizontal axis denotes the ability, but it is also the axis for b. Any item in a test provides some information about the ability of the examinee. The amount of this information depends on how closely the difficulty of the item matches the ability of the person. The item information function of the 1PL model is computed as follows:

$$I_j(\theta, b_j) = P_j(\theta, b_j)Q_j(\theta, b_j) \tag{3.25}$$

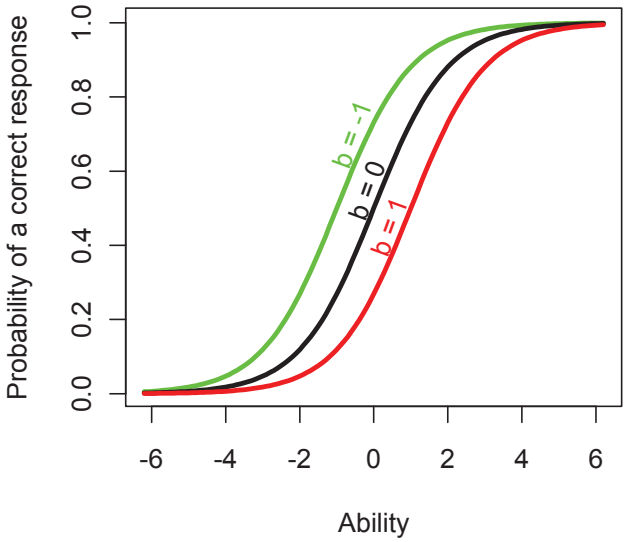


Figure 3.1: Item characteristic curves for the 1PL model

The maximum value of the item information function is 0,25 when the probabilities of a correct and of an incorrect response are equal to 0,5.
The test information function (TIF) is the sum of the item information functions.

$$I_i(\theta_i) = \sum_j I_{ij}(\theta_i, b_j) \tag{3.26}$$

where the index i refers to the person, and the index j refers to the item.
The variance of the ability estimate $\hat{\theta}$ can be estimated as the reciprocal value of the test information function at $\hat{\theta}$. The standard error of measurement (SEM) is equal to the square root of the variance.

$$Var(\hat{\theta}) = \frac{1}{I(\hat{\theta})} \tag{3.27}$$

Two parameter logistic item response function

The ICCs of all items are not always parallel. So we need another model that can fit the data better. The 2PL model allows for different slopes called the item´s discrimination a. Three 2PL ICCs have been drawn for items with the same parameter b (b = 0) and three different slopes a ($a = 0.5, 1, 2$) in Figure 3.28. The items with large slopes are easier to discriminate between lower and higher proficiency examinees.

$$P_{ij}(\theta_i, b_j, a_j) = \frac{exp[a_j(\theta_i - b_j)]}{1 + exp[a_j(\theta_i - b_j)]} \tag{3.28}$$

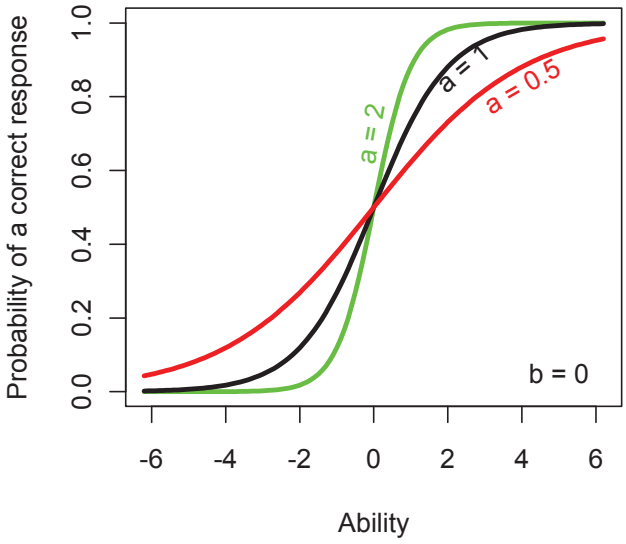


Figure 3.2: Item characteristic curves for the 2PL model

Ayearst and Bagby (2011) recommend that the discrimination coefficients smaller than 0,40 provide less information for estimation of θ. According to Ayala (2009), discrimination parameters of items should range in an interval [0.8, 2.5].

The item information function of the 2PL model is defined as follows

$$I_j(\theta, b_j, a_j) = a_j^2 P_j(\theta; b_j) Q_j(\theta; b_j) \tag{3.29}$$

The item information will increase substantially as discrimination parameters above one and vice versa because it appears in the formula as a square. In the 2PL model, the item information functions still obtain the maxima at item difficulty like in the 1PL model. However, the values of the maxima depend on the discrimination parameter. Items with high discrimination parameters are most informative and the information is concentrated around item difficulty. The information will spread along the ability axis with smaller values when a is low.

The test information function of the 2PL model is the sum of the item information functions over the items in a test as follows. The variance of the ability estimate and SEM in the 2PL model are calculated as in the 1PL model.

$$I_i(\theta_i) = \sum_j I_{ij}(\theta_i, b_j, a_j) \tag{3.30}$$

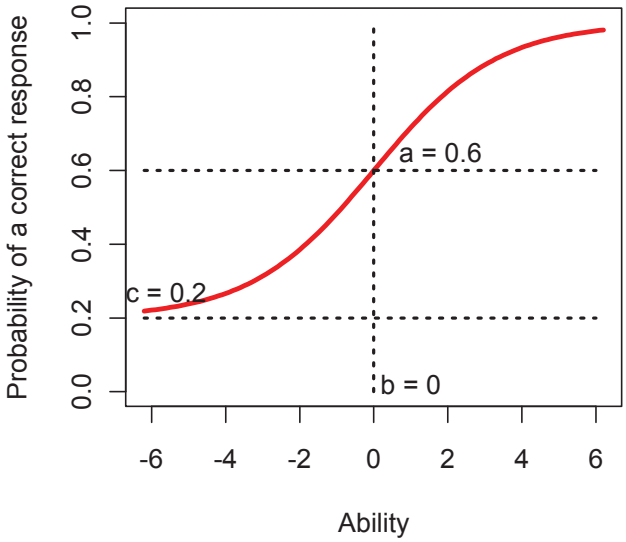


Figure 3.3: Item characteristic curve for the 3PL model

Three parameter logistic item response function

Neither of these two above models allows for a guessing of the multiple-choice item. Allan Birnbaum (1968) developed a new model with a guessing parameter c. Figure 3.3 depicts an example of the 3PL model with a=0,6, b=0, c=0,2. The 3PL model is the most commonly applied IRT model in testing assessments with the following formula. We just utilize 1PL and 2PL models in this work.

$$P(\theta, a, b, c) = c + (1 - c)\frac{exp[a(\theta - b)]}{1 + exp[a(\theta - b)]} \tag{3.31}$$

3.4.3 Estimating proficiency

In this section, we will present how to estimate proficiency of 1PL and 2PL models. One of the most important assumptions in IRT is local independence within items. This means that the responses given to the items in a test are mutually independent. Therefore, we can multiply probability of each response given ability to obtain the probability of the whole pattern.

Assume that the parameter β (b for 1PL model and a, b for 2PL model) for each item have been already estimated. The method of maximum likelihood estimation will be introduced. The likelihood function is shown in the following formula.

$$L(\theta) = \prod_j P_j(\theta, b_j)^{x_j} Q_j(\theta, b_j)^{1-x_j} \tag{3.32}$$

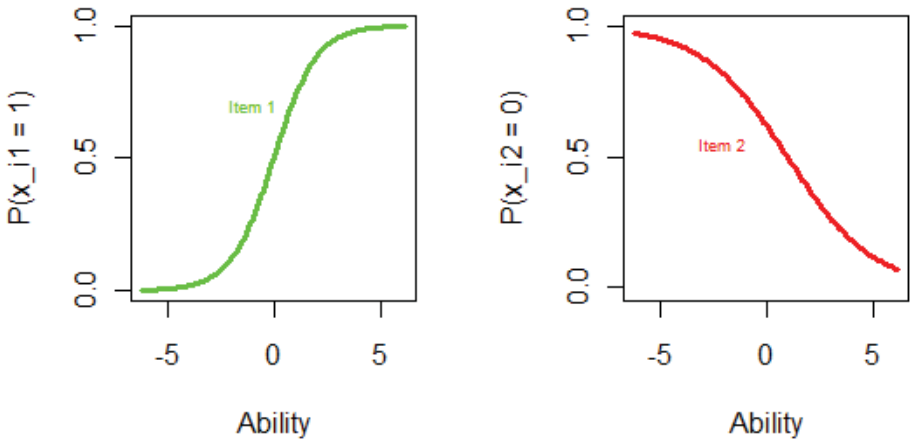


Figure 3.4: ICCs of item 1 (correct response), item 2 (incorrect response)

where $x_j \in (0, 1)$ is the score on item j. The likelihood is used to predict latent ability from the observed responses. The ability $\hat{\theta}$, which has the highest likelihood given the item parameters, will become the ability estimate.

The first term $P(\theta)$ in the equation 3.32 reflects the ICCs for correct responses, the second term $Q(\theta)$ for incorrect responses. Figure 3.4 shows the probabilities for each example item. Figure 3.5 shows the likelihood for two items.

3.4.4 Estimating item parameters

We have assumed that the item parameters were known. This is never the case in practice. The 1PL and 2PL models are utilized to fit the statistics exams. The probabilities of the responses are given as

$$P(X|\theta, \beta) = \prod_i P(x_i|\theta_i, \beta) = \prod_i \prod_j P(x_{ij}|\theta_i, \beta_j) \tag{3.33}$$

where $\theta = (\theta_1, ..., \theta_n)$, $\beta = (a_1, b_1, ..., a_m, b_m)$ are unknown, fixed parameters

Let $p(\theta)$ represent prior knowledge about the examinee distribution. This distribution of latent abilities is assumed a standard normal distribution. Maximum marginal likelihood (MML) estimates of β maximize the following function.

$$L(\beta|X) = \prod_i \int p(x_i|\theta, \beta)p(\theta)d\theta \tag{3.34}$$

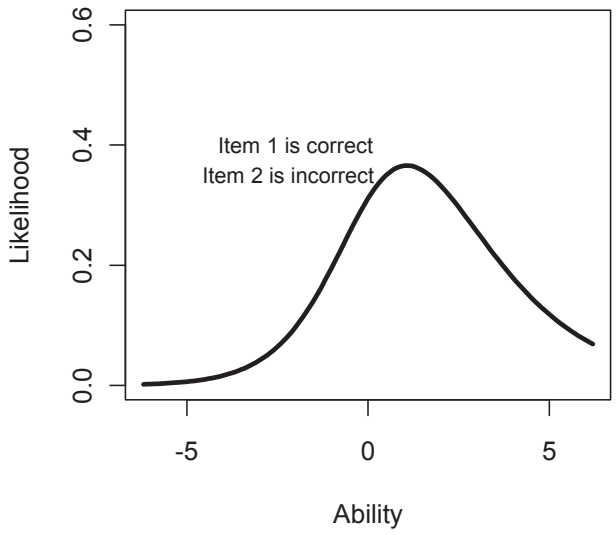


Figure 3.5: ICCs multiply together to yield the likelihood

3.4.5 Goodness of fit indices

Goodness of fit indices for 1PL IRT model

In this part two goodness of fit indices for 1PL IRT model are introduced. All computations were implemented using eRm (extended Rasch modeling) package in R.

The first goodness-of-fit approach is based on the observed responses $x_j \in (0, 1)$ and the estimated model probabilities $\hat{P}_{ij}$. The deviance is formed as follows:

$$D_0 = -2\sum_{i=1}^{n}\sum_{j=1}^{m}\left[x_{ij}log\left(\frac{\hat{P}_{ij}}{x_{ij}}\right) + (1 - x_{ij})log\left(\frac{1-\hat{P}_{ij}}{1-x_{ij}}\right)\right] \tag{3.35}$$

where the index i refers to the person, and the index j refers to the item. We represent the matrix X as a vector x of length $l = 1, ..., L$ where $L = n \times m$. D_0 is computed with another formula using the deviance residuals due to possible 0 values in the denominators in the 3.35.

$$D_0 = \sum_{l=1}^{L} d_l^2 \tag{3.36}$$

$$d_l = \sqrt{2\left|log(\hat{P}_l)\right|} \qquad \forall x_l = 1$$

$$d_l = \sqrt{-2\left|log(1-\hat{P}_l)\right|} \qquad \forall x_l = 0$$

The second goodness-of-fit measure is McFadden´s R^2 (McFadden 1974), which can be expressed as

$$R^2_{MF} = \frac{LogL_0 - LogL_G}{LogL_G} \tag{3.37}$$

where L_G is the likelihood of any IRT model, L_0 is the likelihood of intercept-only model, in which there are neither item effects nor person effects. It is explained as proportional reduction in the deviance statistic (Menard 2000).

Goodness of fit indices for 1PL and 2PL IRT model

The approach used primarily for model comparisons is information criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC). For competing models, the model which minimizes an information criteria is normally selected.

$$AIC = -2lnL + 2n \tag{3.38}$$

$$BIC = -2lnL + ln(m)n \tag{3.39}$$

where n is the number of estimated parameters, m is the number of persons.

3.5 Polytomous Item Response Theory

This chapter will be devoted to a polytomous IRT model for testlets. A testlet is a group of locally-dependent items. Tests may consist of several testlets and seperate independent items. A reading passage or a graph with some related questions is considered as a testlet. One of the shortcomings of dichotomous IRT is the assumption of local independence of items. The questions of several exercises in both statistics exams seem to be not independent with each other. In order to make the local dependencies within testlets disappear, we can consider the entire testlet (exercise) as a unit and score it polytomously (Wainer, 2007).

3.5.1 Model Specification

The polytomous IRT model presented here is graded response model (Samejima 1969). The definition and explanation were taken from Mark D. Reckase, 2009 and Wainer, 2007. The applied parameters estimation approach is marginal maximum likelihood (MML) approach.

The characteristics of graded response (GR) model is the successful accomplishment of one step requires the successful accomplishment of the previous steps. The probability of accomplishing k or more steps is assumed to increase monotonically with an increasing θ. The probability of receiving a score of k also called category response function is:

$$P(u_{ij} = k|\theta_j) = P^*(u_{ij} = k|\theta_j) - P^*(u_{ij} = k+1|\theta_j) \tag{3.40}$$

with

$$P^*(u_{ij} = k|\theta_j) = \frac{e^{a_i(\theta_j - b_{ik})}}{1 + e^{a_i(\theta_j - b_{ik})}} \tag{3.41}$$

where k is the score on the item ($k = 0, 1, ..., m_i$), $P^*(u_{ij} = 0|\theta_j) = 1$, $P^*(u_{ij} = m_i + 1|\theta_j) = 0$ and $P^*(u_{ij} = k|\theta_j)$ is the cumulative category response function of k steps.

The logistic form of the GR model is shown as follows:

$$P(u_{ij} = k|\theta_j) = \frac{e^{a_i(\theta_j - b_{ik})} - e^{a_i(\theta_j - b_{i,k+1})}}{(1 + e^{a_i(\theta_j - b_{ik})})(1 + e^{a_i(\theta_j - b_{i,k+1})})} \tag{3.42}$$

where a_i is an item discrimination parameter, b_{ik} is a difficulty parameter for the k-th step of the item.

3.5.2 Expected score

The expected score of an item in the GR model is the sum of the products of the probability of an item score and the item score and is calculated with the following formula

$$E(u_{ij}|\theta_j) = \sum_{k=0}^{m_i} kP(u_{ij} = k|\theta_j) \tag{3.43}$$

where k is the score on the item ($k = 0, 1, ..., m_i$). The expected score of polytomous items is interpreted similarly as the ICC of dichotomously scored items.

3.5.3 Reliability

Reliability measures the precision of tests and can be characterized as a function of proficiency θ. The marginal reliability is

$$\overline{\rho} = \frac{\sigma_\theta^2 - \overline{\sigma}_{e^*}^2}{\sigma_\theta^2} \tag{3.44}$$

$$\overline{\sigma}_{e^*}^2 = \int \sigma_{e^*}^2 g(\theta) d\theta \tag{3.45}$$

where $\overline{\sigma}_{e^*}^2$ is the marginal error variance, $g(\theta)$ is proficiency density fixed as N(0,1), $\sigma_{e^*}^2$ is the expected value of the error variance. The error variance function can be calculated from the information function I(θ). The integration in 3.45 could be implemented and $\bar{\rho}$ is calculated through 3.44.

3.5.4 Information function

The information function (IF) for the GR model derived by Samejima (1969) is given in the following equation. The IF points out how many standard errors of the trait estimate are needed to equal one unit on the proficiency scale. When more standard errors are needed to equal one unit on the θ-scale, the standard errors are smaller indicating that the measuring instrument is sensitive enough to detect relatively small differences in θ (Lord, 1980 and Hambleton & Swaminathan, 1985).

The information for a scale item in the GR model is a weighted sum of the information from each of the response alternatives. If the information plot is not

unimodal it will affect the selection of items e.g. in adaptive tests to maximize the information.

$$I(\theta_j, u_i) = \sum_{k=1}^{m+1} \frac{\left[a_i P^*_{i,k-1}(\theta_j) Q^*_{i,k-1}(\theta_j) - a_i P^*_{ik}(\theta_j) Q^*_{ik}(\theta_j)\right]^2}{P^*_{i,k-1}(\theta_j) - P^*_{ik}(\theta_j)} \tag{3.46}$$

where $P^*_{ik}(\theta_j) = P^*(u_{ij} = k)|\theta_j)$ is the cumulative category response function of k steps of the item and $Q^*_{ik}(\theta_j) = 1 - P^*_{ik}(\theta_j)$.

Chapter 4

IRTPRO 2.1 for Windows

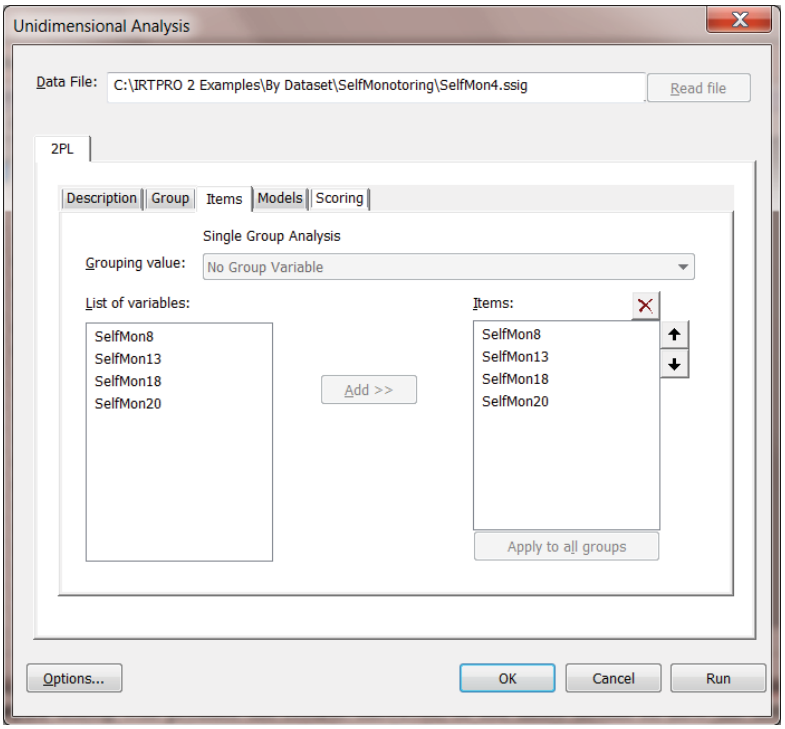
IRTPRO (Item Response Theory for Patient-Reported Outcomes) is a statistical software for item calibration and test scoring using IRT. IRTPRO 2.1 for Windows is developed by Li Cai, David Thissen & Stephen du Toit. This product has replaced the four programs Bilog-MG, Multilog, Parscale and Testfact. The program has been tested on the Microsoft Windows platform with Windows7, Vista and XP operating systems.
Various IRT models can be implemented in IRTPRO, for example:

1. Two-parameter logistic model(2PL)
2. Three-parameter logistic model(3PL)
3. Graded model
4. Generalized Partial Credit model
5. Nominal model

IRTPRO implements the method of maximum likelihood for item parameter estimation, or it computes maximum a posteriori (MAP) estimates if prior distributions are specified for the item parameters.
IRT scores in IRTPRO can be computed using any of the following methods:

- Maximum a posteriori (MAP) for response patterns
- Expected a posteriori (EAP) for response patterns
- Expected a posteriori (EAP) for summed scores

IRTPRO supports both model-based and data-based graphical displays. Figure 4.1 shows two examples of graphical displays. More information about software IRTPRO 2.1 under http://www.ssicentral.com.



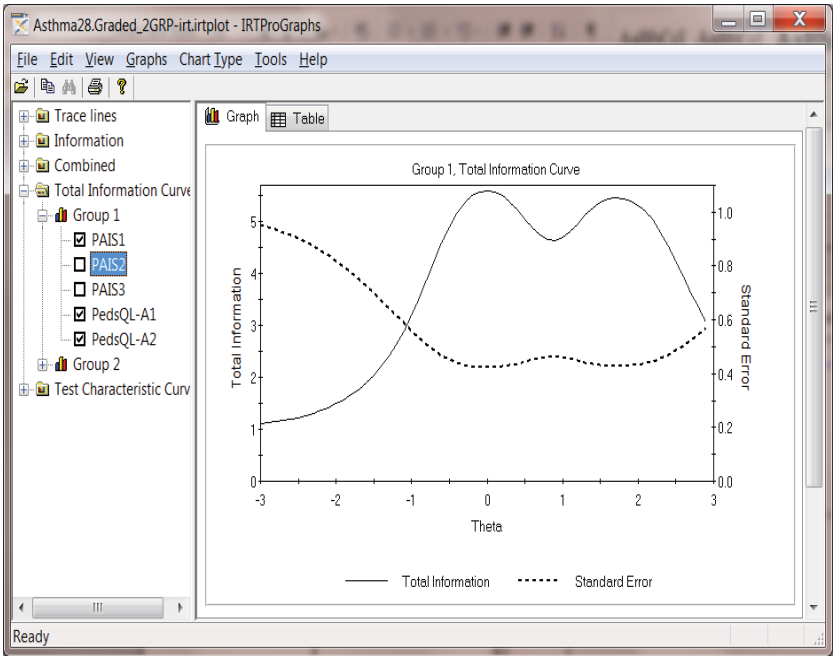


Figure 4.1: Graphical examples of software IRTPRO 2.1

Chapter 5

Reliability analysis

In this section, the coefficients for reliability analysis were computed for each data set with the formula 3.1 and 3.2.

- Cronbach´s alpha coefficient α based on covariances
- standardized Cronbach´s alpha coefficient α_{st} based on tetrachoric correlations
- the average inter-item correlation $\bar{r}$

where n is the number of observations, m is the number of items in scale.

Table 5.1 showed the Cronbach´s α coefficients in both exams are really high, which means a set of items measures a single unidimensional latent construct (ability of examinees) very well. The standardized Cronbach´s α of the first exam is larger than that of the second exam. Hence, we could conclude that the items in the first scale made a better measurement of the ability of examinees.

The values of Cronbach´s α have not increased when any of the items was dropped in the first data set. So all questions should be kept for this test. If we drop the third question of exercise 3 or the last question of exercise 6 in the second test, Cronbach´s α will increase by 1% to $\alpha_{st} = 0,94$. The third question of exercise 3 is about the calculation of a univariate variable. Theoretical distribution function is the content of the last question of exercise 6. In comparison to the other items these two items correlate less with the entire scale possessing the values 0,21 and 0,29, respectively. For this reason, the contribution of both items to the exam should be reconsidered.

Exam	n	m	α	α_{st}	$\bar{r}$
1	176	28	0,90	0,95	0,41
2	171	23	0,86	0,93	0,38

Table 5.1: Reliability analysis of the two exams

Chapter 6

Exploratory factor analysis

6.1 Tetrachoric correlation

Figure 6.1 and 6.2 depict the tetrachoric correlations of 28 questions and 23 questions in the first and second exams. Assuming that there are latent continuous variables underlying the dichotomized variables which are normally distributed, the tetrachoric correlation estimates the correlation between the assumed underlying continuous variables. The items in these figures are somehow correlated. Lightly-pink color squares indicated negative correlation between items. The EFA is performed based on the correlation matrix. How tetrachoric correlations are computed? See section 3.2.1 for more details.

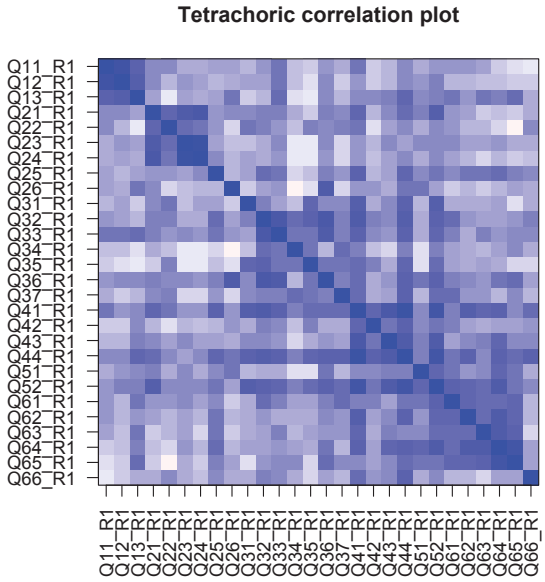


Figure 6.1: Tetrachoric correlation of 28 questions in the first exam

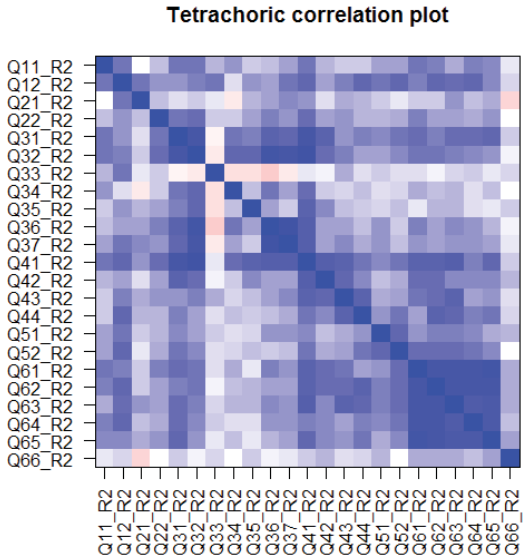


Figure 6.2: Tetrachoric correlation of 23 questions in the second exam

Exam	Kaiser	Horn
1	6	4
2	6	3

Table 6.1: Number of extracted factors according to Kaiser and Horn´s criterion

6.2 Estimation of factor model

As mentioned earlier, EFA is applied to explore the interrelationships between the observed variables and to determine whether the interrelationships explained by a small number of latent variables. We applied the principal component method based on tetrachoric correlation matrix. Varimax rotation was taken into account to achieve a better interpretation of factor loadings. The estimation method is weighted least squares for binary data. All computations were done with software Mplus and R.

According to the Kaiser criterion, factors should be extracted when their eigenvalues are bigger than one. An eigenvalue of a factor indicates variance of all variables which is explained by the factor.

Horn's parallel analysis is a Monte-Carlo based simulation method that compares the observed eigenvalues with those obtained from uncorrelated normal variables. A factor is retained if the associated eigenvalue is bigger than the 95th of the distribution of eigenvalues derived from the random data (Wikipedia). This method is one of the most recommendable rules for determining the number of factors. The number of extracted factors according to Kaiser criterion and Horn´s parallel analysis in the first and second exams was shown in the Table 6.1.

No.	Question	Theory/Practice	Field	F1	F2
1	Q11_R1	Theory	Combinatorics	0,79	
2	Q12_R1	Theory	Combinatorics	0,80	
3	Q13_R1	Theory	Combinatorics	0,47	
4	Q21_R1	Theory	Probability	0,71	
5	Q22_R1	Theory	Probability	0,53	
6	Q23_R1	Theory	Probability	0,92	
7	Q24_R1	Theory	Probability	0,94	
8	Q25_R1	Theory	Probability	0,55	
9	Q26_R1	Theory	Probability	0,62	
10	Q31_R1	Practice	Univariate		0,52
11	Q32_R1	Practice	Univariate		0,68
12	Q33_R1	Practice	Univariate		0,59
13	Q34_R1	Practice	Univariate		0,52
14	Q35_R1	Practice	Univariate		0,59
15	Q36_R1	Practice	Univariate		0,59
16	Q37_R1	Practice	Univariate		0,60
17	Q41_R1	Practice	Bivariate		0,91
18	Q42_R1	Practice	Bivariate		0,55
19	Q43_R1	Practice	Bivariate		0,73
20	Q44_R1	Practice	Bivariate		0,89
21	Q51_R1	Practice	Bivariate		0,50
22	Q52_R1	Practice	Bivariate		0,77
23	Q61_R1	Theory	Univariate		0,62
24	Q62_R1	Theory	Univariate		0,71
25	Q63_R1	Theory	Univariate		0,69
26	Q64_R1	Theory	Univariate		0,82
27	Q65_R1	Theory	Univariate		0,85
28	Q66_R1	Theory	Distribution		0,55

Table 6.2: Two-factor model of the first exam

6.2.1 Factor model for the first exam

Based on Kaiser criterion six factors should be extracted. This criterion often tends to overextract factors. In six-factor model, several factors which have not significant loadings are incomprehensible. Hence, the six-factor model was not introduced here. Horn´s parallel analysis determined to extract four factors. For those reasons two-, three-, four-factor models were conducted for the first exam.

Performing two-factor model analysis for the 28 items in the first exam, we achieved the factor loadings presented in Table 6.2. Nine questions of exercises 1 and 2 are loaded on the first factor. We can call this factor combinatorics-probability factor. The exercises 3, 4, 5 and 6 are loaded on the second factor named univariate-bivariate factor. Clearly, EFA has shown the relationships between the related questions with respect to the content of questions.

EFA was conducted again with three- and four-factor model. The results were exhibited in Table 6.3 and 6.4. In three-factor model the exercises 1 and 2 are also loaded on the first factor. The second factor has strong loadings on

No.	Question	Theory/Practice	Field	F1	F2	F3
1	Q11_R1	Theory	Combinatorics	0,62		
2	Q12_R1	Theory	Combinatorics	0,66		
3	Q13_R1	Theory	Combinatorics	0,45		
4	Q21_R1	Theory	Probability	0,68		
5	Q22_R1	Theory	Probability	0,46		
6	Q23_R1	Theory	Probability	0,93		
7	Q24_R1	Theory	Probability	0,88		
8	Q25_R1	Theory	Probability			0,52
9	Q26_R1	Theory	Probability		0,68	
10	Q31_R1	Practice	Univariate		0,48	
11	Q32_R1	Practice	Univariate		0,72	
12	Q33_R1	Practice	Univariate		0,65	
13	Q34_R1	Practice	Univariate		0,43	
14	Q35_R1	Practice	Univariate		0,62	
15	Q36_R1	Practice	Univariate		0,82	
16	Q37_R1	Practice	Univariate		0,49	
17	Q41_R1	Practice	Bivariate			0,77
18	Q42_R1	Practice	Bivariate			0,46
19	Q43_R1	Practice	Bivariate			0,70
20	Q44_R1	Practice	Bivariate			0,70
21	Q51_R1	Practice	Bivariate			0,48
22	Q52_R1	Practice	Bivariate			0,67
23	Q61_R1	Theory	Univariate			0,59
24	Q62_R1	Theory	Univariate			0,70
25	Q63_R1	Theory	Univariate			0,69
26	Q64_R1	Theory	Univariate			0,89
27	Q65_R1	Theory	Univariate			0,86
28	Q66_R1	Theory	Distribution			0,45

Table 6.3: Three-factor model of the first exam

seven questions of exercise 3. All these questions are about the calculation of univariate variable. The last three exercises whose content expresses the practice of bivariate variables and theoretical univariate variable are loaded on the third factor.

Table 6.5 depicts eigenvalues and proportions of explained variance in the first exam. Six eigenvalues are larger than one which make an explanation of 78% of total variance. Only the first eigenvalue has already explained 46% variance of all 28 questions. Each of the last five eigenvalues explained less than 10%.

6.2.2 Factor model for the second exam

According to Kaiser criterion also six factors should be extracted for the second exam. Extraction of three factors is the determination of Horn´s parallel analysis. Loadings of variables on the three-, four-, five-factor models are not very interesting. Hence, only the figures of two-factor model were presented here.

No.	Question	Theory/Practice	Field	F1	F2	F3	F4
1	Q11_R1	Theory	Combinatorics	0,81			
2	Q12_R1	Theory	Combinatorics	0,85			
3	Q13_R1	Theory	Combinatorics	0,60			
4	Q21_R1	Theory	Probability				0,69
5	Q22_R1	Theory	Probability				0,58
6	Q23_R1	Theory	Probability				0,91
7	Q24_R1	Theory	Probability				0,92
8	Q25_R1	Theory	Probability			0,47	
9	Q26_R1	Theory	Probability	0,75			
10	Q31_R1	Practice	Univariate		0,63		
11	Q32_R1	Practice	Univariate		0,75		
12	Q33_R1	Practice	Univariate		0,56		
13	Q34_R1	Practice	Univariate		0,56		
14	Q35_R1	Practice	Univariate		0,82		
15	Q36_R1	Practice	Univariate	0,73			
16	Q37_R1	Practice	Univariate		0,60		
17	Q41_R1	Practice	Bivariate		0,74		
18	Q42_R1	Practice	Bivariate		0,41		
19	Q43_R1	Practice	Bivariate			0,58	
20	Q44_R1	Practice	Bivariate		0,72		
21	Q51_R1	Practice	Bivariate			0,52	
22	Q52_R1	Practice	Bivariate			0,54	
23	Q61_R1	Theory	Univariate			0,55	
24	Q62_R1	Theory	Univariate			0,68	
25	Q63_R1	Theory	Univariate			0,69	
26	Q64_R1	Theory	Univariate			0,90	
27	Q65_R1	Theory	Univariate			0,90	
28	Q66_R1	Theory	Distribution		0,46		

Table 6.4: Four-factor model of the first exam

Eigenvalue	Proportion of variance	Cumulated proportion
12,8	0,46	0,46
2,4	0,09	0,55
2,3	0,08	0,63
1,7	0,06	0,69
1,4	0,05	0,74
1,2	0,04	0,78

Table 6.5: Eigenvalues and proportions of explained variance in the first exam

Table 6.6 presented the two-factor model of the 23 questions in the second exam. Almost questions of exercises 1, 4, 5 and 6 are loaded on the first factor. There are much more theoretical questions in the first factor. The second factor has strong loadings on nearly all questions of exercises 2 and 3 whose content is more practical. We can name the second factor as the practical factor.

The loadings of exercises 1, 2 as well as 3 are not concentrated on one single

No.	Question	Theory/Practice	Field	F1	F2
1	Q11_R2	Theory	Combinatorics	0,43	
2	Q12_R2	Theory	Probability	0,64	
3	Q21_R2	Theory	Probability		0,31
4	Q22_R2	Theory	Probability		0,46
5	Q31_R2	Practice	Univariate		0,71
6	Q32_R2	Practice	Univariate		0,93
7	Q33_R2	Practice	Univariate	0,32	
8	Q34_R2	Practice	Univariate		0,59
9	Q35_R2	Practice	Univariate		0,56
10	Q36_R2	Practice	Univariate		0,92
11	Q37_R2	Practice	Univariate		0,91
12	Q41_R2	Practice	Bivariate		0,80
13	Q42_R2	Practice	Bivariate	0,48	
14	Q43_R2	Practice	Bivariate	0,57	
15	Q44_R2	Practice	Bivariate	0,61	
16	Q51_R2	Theory	Bivariate	0,49	
17	Q52_R2	Theory	Bivariate	0,60	
18	Q61_R2	Theory	Univariate	0,87	
19	Q62_R2	Theory	Univariate	0,89	
20	Q63_R2	Theory	Univariate	0,80	
21	Q64_R2	Theory	Univariate	0,83	
22	Q65_R2	Theory	Univariate	0,86	
23	Q66_R2	Theory	Distribution	0,34	

Table 6.6: Two-factor model of the second exam

Eigenvalue	Proportion of variance	Cumulated proportion
10,36	0,45	0,45
2,54	0,11	0,56
1,64	0,07	0,63
1,34	0,06	0,69
1,28	0,06	0,75
1,05	0,05	0,80

Table 6.7: Eigenvalues and proportions of explained variance in the second exam

factor in the four-factor model. The first factor loaded all questions in the exercise 6 with relatively high loadings. We can let the second factor load the exercises 1, 2 and 5. The third factor has strong loadings on most items of exercise 3. Exercise 4 is loaded on the fourth factor.

The five-factor model is not much different as the four-factor model. The exercises 2, 3, 4 and 6 are loaded on separate factor. The exercises 1 and 5 are loaded on the same factor. Table 6.7 showed eigenvalues and proportions of explained variance in the second exam. 45% of total variance was explained through the first factor. All six factors would explain 80% of the variation of all 23 items. Each of the last four eigenvalues made an explanation less than 10%.

Chapter 7

Confirmatory factor analysis

As referred earlier, CFA is used to test whether the data fit a hypothesized measurement model proposed by a researcher. This hypothesized model is based on theory or previous study. The difference to EFA is that each variable is just loaded on one factor. Error terms contain the remaining influence of variables.

In this section, we would consider the two-factor model of the first and second exams. In the first exam we achieved two factors from EFA, called the combinatorics-probability factor and the univariate-bivariate factor. The two factors in the second exam are named theoretical and practical factor. If the two factors in two exams are valid, then all questions belonging to the factor should measure a single construct. The evaluation of unidimensionality of questions in each factor would provide information about the validity of these contructs. All computations for CFA were done using Mplus.

The Chi-squared test checks the hypothesis that the theoretical covariance matrix corresponds to the empirical covariance matrix. The test statistic is χ^2-distributed under the assumption of the null hypothesis. The null hypothesis will be rejected if the value of the test statistic is large. Due to a lack of statistical power from small sample sizes, one may fail to reject the hypothesis (or accept the model). That is the type I error. Likewise, type II error will occur when large samples are used since big n will make the test always significant.

The null hypothesis of all various models in both exams was rejected due to the large sample sizes. Hence, other measures of fit have been taken into account.

Goodness of fit indices such as RMSEA, TLI and CFI were used to evaluate to what extent a particular factor model explains the empirical data (Muthén, 2004). See section 3.3.2 for more details about the indices.

Table 7.1 and 7.2 depicted the values of RMSEA, TLI and CFI of the first and second exams with respect to various factor models. RMSEA smaller than 0,05 will indicate a good model fit. TLI and CFI larger than 0,95 are indicative of good model fit. The values RMSEA, TLI and CFI of the first exam are unacceptable for two-factor model. Five-factor model is a good model based on the RMSEA, TLI and CFI. The two-factor model explained the empirical data of the second exam very well according to these goodness of fit indices. Hence,

Factor model	$RMSEA_1$	CFI_1	TLI_1
Two-factor	0,073	0,91	0,91
Three-factor	0,070	0,92	0,92
Four-factor	0,063	0,94	0,93
Five-factor	0,050	0,96	0,95

Table 7.1: Goodness model of fit in the first exam

Factor model	$RMSEA_2$	CFI_2	TLI_2
Two-factor	0,05	0,96	0,95
Three-factor	0,04	0,97	0,96
Four-factor	0,04	0,98	0,97
Five-factor	0,04	0,98	0,97

Table 7.2: Goodness model of fit in the second exam

we can conclude that exercises 1, 4, 5, and 6 in the second exam measure a single contruct - the theoretical factor. Exercises 2 and 3 measure another contruct - the practical factor.

Chapter 8

Dichotomous Item Response Theory

8.1 One parameter logistic item response function

In this section, the analysis of 1PL IRT was indicated. All computations were done using the software R and IRTPRO 2.1. The parameter estimation method is the marginal maximum likelihood method (MML). The simplest IRT identifies each item in term of a single parameter. This parameter is item´s location or difficulty parameter on the latent continuum that represents the contruct. If one wants to measure examinees with different abilities, one needs items of different difficulty level.

The ICCs of all items in 1PL model are always parallel. As mentioned before, the horizontal axis denotes the ability, but it is also the axis for b. Any item in a test provides some information about the ability of the examinee. The amount of this information depends on how closely the difficulty of the item matches the ability of the person. The difficulty parameters b of 28 as well as of 23 questions in the first and second exams are sorted increasingly in Figure 8.1 and 8.2.

In the lower panel of Figure 8.1, we can see that the difficulty parameters of items spread over the range of ability which means examinees of all level abilities measured well by items with all levels of difficulty. The first question in the exercise 4 seems to be the easiest question with smallest $b = -2,43$. The exercise 6 is the most difficult exercise in the test with 4 questions at the end of the range.

In the upper pannel, the upper end of the continuum indicates greater proficiency than does the lower end. Items located toward the right side require an individual to have higher ability to respond items correctly than items toward the left side.

The first statistics exam is a good measurement since examinees with all level of abilities can be measured with the questions of different difficulty along the whole scale. In Figure 8.2, the first question of exercise 3 is the easiest with smallest b. The exercise 6 is the most difficult one with 4 questions standing seemly back of the order. Several more difficult questions are necessary here to

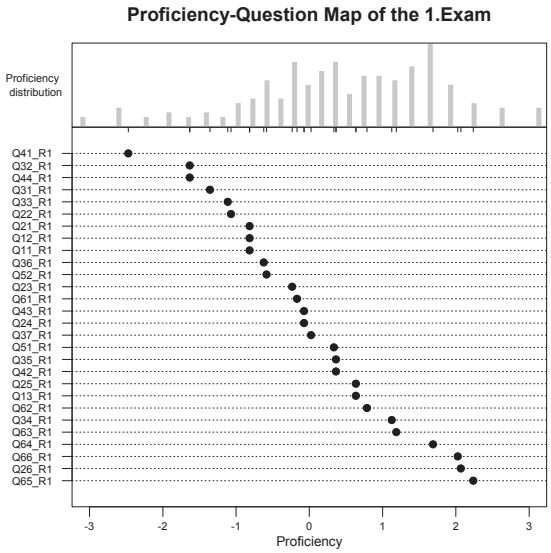


Figure 8.1: Proficiency-Question Map with 28 questions in the first exam

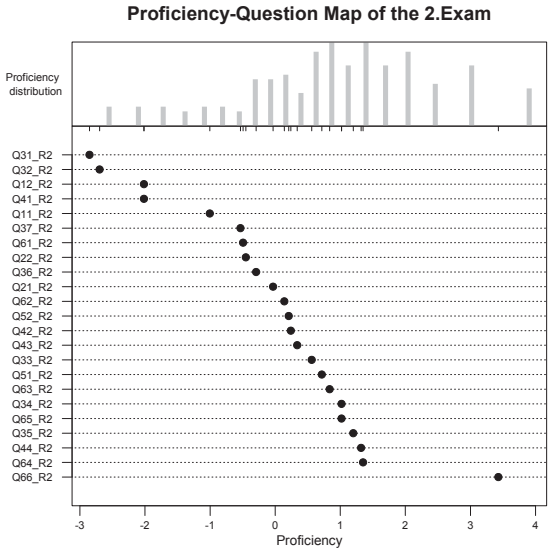


Figure 8.2: Proficiency-Question Map with 23 questions in the second exam

measure high-ability examinees. Maybe it is the purpose of testers to let more examinees pass the second exam.

In the 1PL model, an item gives the most information about examinees when the ability of those examinees is equal to the difficulty of the item. As individual ability becomes either smaller or greater than item difficulty, the

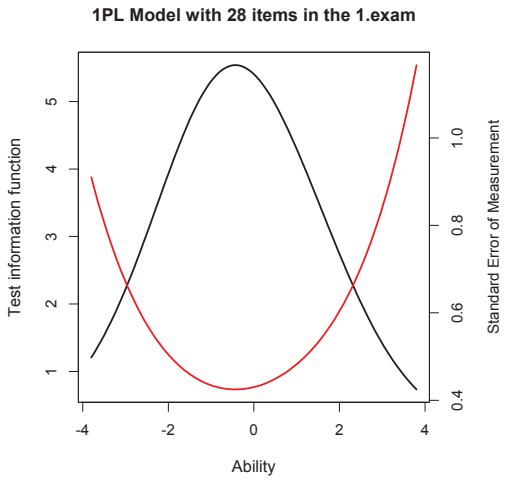


Figure 8.3: TIF and SEM of 28 questions in the first exam

item information decreases. For example, question Q41_R1 yields its maximum information for estimating θ at its difficulty of -2,47 (Table 10.1). Hence, using this item to estimate persons with big θ would not provide precise estimates and the SE would be large.

The test information function (TIF) is calculated as the sum of all item information functions from the scale with the formula3.26. The variance of the ability estimate can be estimated as the reciprocal value of TIF 3.27. The SEM is equal to the square root of the variance.

Figure 8.3 and 8.4 showed the TIF as well as the SEM for 28 items and 23 items in the first and second exam. The TIF is displayed in black, and its values can be read from the left-hand axis. The SEM is displayed in red, and its values can be read from the right-hand axis. The TIF indicates how well the entire instrument can estimate the proficiency θ. The maximum information gained in the first and second exams were around -0,3 and -0,9. The instruments provide less information for large θ. At little left-skewed ability values we achieved the most information and the smallest standard error of two exams.

8.2 Two parameter logistic item response function

The ICCs of all items are not always parallel. The 2PL model allows for different slopes a. Difficulty parameter b of an item is the point on the latent scale where a person has a 50% chance of responding correctly a question. High-difficulty items are less often answered exactly. Items with large slopes are easier to discriminate between lower and higher proficiency examinees.

A comparison of difficulty and discrimination parameters of items would provide the information about the contribution to measuring of the latent trait. Table 8.1 showed the 2PL model item parameter estimates a, b for the first exam. The discrimination coefficients smaller than 0,40 provide less information for estimation of θ and larger than 4 indicate that the assumption of local

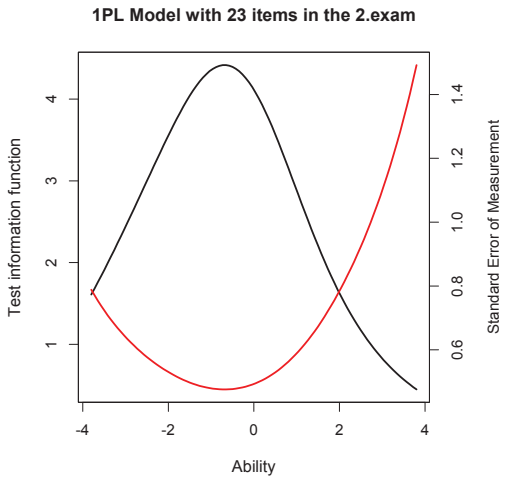


Figure 8.4: TIF and SEM of 23 questions in the second exam

independence is violated (Ayearst and Bagby 2011). According to Ayala (2009), discrimination parameters of items should range in an interval [0,8-2,5]. The slope parameter estimates of the first exam vary from 0,98 (Q34_R1) up to 3,73 (Q52_R1). Five of 28 values are larger than 2,5. In short, most items in the first exam with reasonable values of slope parameters have high enough the separation power.

Table 8.2 contains the 2PL model item parameter estimates a, b for the second exam. Two values of a in exercise 6 exceeding 4 are a signal of violating the assumption of local independence. The slope parameter of $Q41\_R2$ is 2,77 little larger than acceptable value which has loaded on different factor in EFA with respect to the other questions of exercise 4. The discrimination parameter of item $Q33\_R2$ is 0,29, much smaller than acceptable value which can provide less information for estimation of θ. This question is also loaded on different factor in four-, five-factor models. All values of discrimination parameters in both exams are posive which imply that the probability of answering an item correctly increases with increasing θ. The a value of an item either too large or too small would affect the loading of the item on the factor. That is one of the relationships between EFA and IRT models.

Since items possessing low discrimination coefficients indicate that items are not unidimensional. All items in the first exam have rather high discriminatory power. The standardized Cronbach´s alpha coefficient of the first exam is 0,95 which means a set of items measures a single unidimensional latent construct (ability of examinees) very well. Several items (Q21, Q33, Q34, Q35, Q66) of the second exam have low discrimination parameters. Hence, the α_{st} for the second exam is 0,90, smaller than that of the first exam. See section Reliability analysis 5. The exercise 3 seems to differentiate between examinees less than other exercises. $Q33\_R2$ is loaded on different factor in various factor models compared to the other items of exercise 3.

Item	Question	a	SE	b	SE
1	Q11_R1	1,06	0,23	-0,99	0,27
2	Q12_R1	1,02	0,23	-1,02	0,28
3	Q13_R1	1,54	0,32	0,2	0,13
4	Q21_R1	1,4	0,28	-0,8	0,2
5	Q22_R1	1,06	0,24	-1,19	0,3
6	Q23_R1	1,21	0,26	-0,46	0,18
7	Q24_R1	1,06	0,24	-0,38	0,19
8	Q25_R1	1,67	0,33	0,2	0,12
9	Q26_R1	1,15	0,31	1,39	0,31
10	Q31_R1	1,23	0,26	-1,26	0,28
11	Q32_R1	1,77	0,35	-1,14	0,22
12	Q33_R1	1,72	0,35	-0,87	0,19
13	Q34_R1	0,98	0,25	0,68	0,21
14	Q35_R1	1,09	0,24	-0,01	0,17
15	Q36_R1	2,01	0,4	-0,52	0,14
16	Q37_R1	1,34	0,28	-0,23	0,15
17	Q41_R1	2,98	0,67	-1,24	0,2
18	Q42_R1	1,26	0,27	0,01	0,15
19	Q43_R1	1,8	0,35	-0,23	0,13
20	Q44_R1	3,43	0,78	-0,83	0,14
21	Q51_R1	1,51	0,31	0	0,13
22	Q52_R1	3,73	0,79	-0,36	0,1
23	Q61_R1	1,79	0,35	-0,29	0,13
24	Q62_R1	2	0,4	0,28	0,11
25	Q63_R1	1,91	0,39	0,52	0,12
26	Q64_R1	3,57	0,8	0,67	0,1
27	Q65_R1	3,02	0,7	0,95	0,14
28	Q66_R1	1,35	0,34	1,22	0,24

Table 8.1: 2PL model item parameter estimates for the first exam

Item	Question	a	SE	b	SE
1	Q11_R2	1,08	0,27	-1,63	0,45
2	Q12_R2	1,6	0,39	-1,91	0,47
3	Q21_R2	0,6	0,2	-1,33	0,5
4	Q22_R2	1,02	0,24	-1,23	0,4
5	Q31_R2	1,99	0,57	0,4	0,54
6	Q32_R2	2,18	0,65	-2,01	0,5
7	Q33_R2	0,29	0,17	-1,05	0,82
8	Q34_R2	0,61	0,19	0,11	0,35
9	Q35_R2	0,61	0,19	0,36	0,36
10	Q36_R2	1,24	0,27	-0,95	0,31
11	Q37_R2	1,3	0,29	-1,1	0,31
12	Q41_R2	2,77	0,95	-1,51	0,43
13	Q42_R2	1,37	0,29	-0,5	0,27
14	Q43_R2	1,54	0,33	-0,39	0,26
15	Q44_R2	1,67	0,34	0,28	0,27
16	Q51_R2	1,13	0,25	-0,17	0,28
17	Q52_R2	1,39	0,32	-0,52	0,26
18	Q61_R2	2,97	1,4	-0,7	0,15
19	Q62_R2	4,45	2,38	-0,33	0,18
20	Q63_R2	4,06	1,03	-0,01	0,22
21	Q64_R2	3,03	0,78	0,24	0,24
22	Q65_R2	3,08	0,8	0,07	0,23
23	Q66_R2	0,55	0,3	3,95	1,96

Table 8.2: 2PL model item parameter estimates for the second exam

Chapter 9

Polytomous Item Response Theory

9.1 Data analysis

In this section, we illustrate how the polytomous IRT model can yield an analysis of test data. The data is from the Statistics Exams for Undergraduate Students of School of Business and Economics bzw. Ladislaus von Bortkiewicz Chair of Statistics, Humboldt University, Berlin in summer semester 2011.

The statistics exams composed of six exercises in the first and second round.

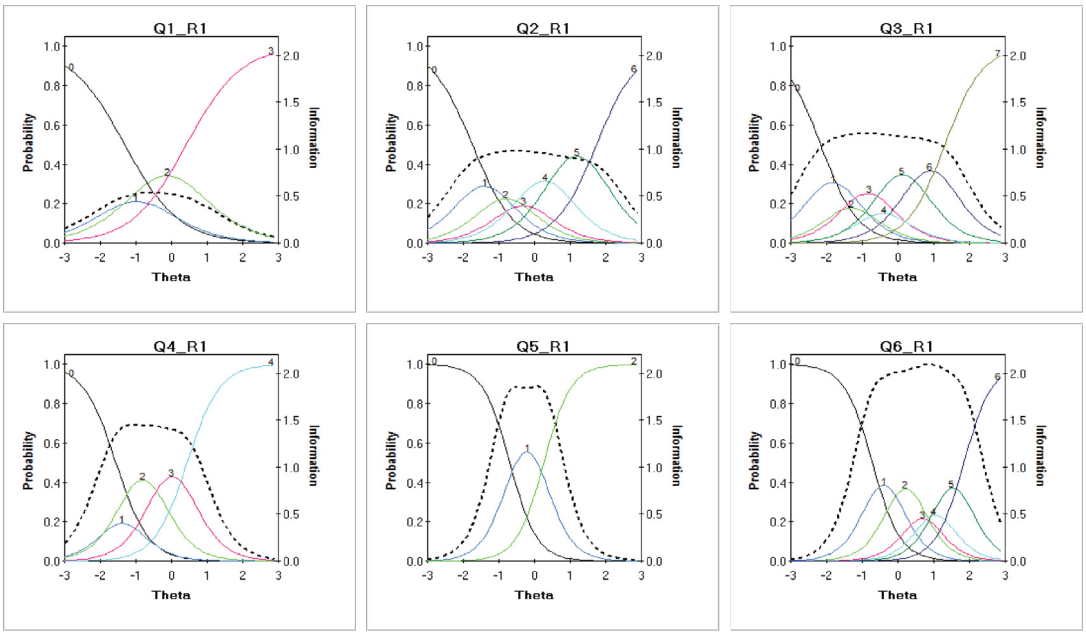


Figure 9.1: Trace lines of 6 exercises in the first exam

Exercise	a^{SE}	b_1	b_2	b_3	b_4	b_5	b_6	b_7
E1_R1	$1,31^{0,2}$	$-1,31^{0,2}$	$-0,66^{0,2}$	$0,42^{0,1}$				
E2_R1	$1,74^{0,3}$	$-1,75^{0,2}$	$-1,07^{0,2}$	$-0,54^{0,1}$	$-0,10^{0,1}$	$0,65^{0,1}$	$1,74^{0,2}$	
E3_R1	$1,90^{0,3}$	$-2,17^{0,3}$	$-1,50^{0,2}$	$-1,12^{0,2}$	$-0,58^{0,1}$	$-0,26^{0,1}$	$0,50^{0,1}$	$1,31^{0,2}$
E4_R1	$2,18^{0,3}$	$-1,56^{0,3}$	$-1,21^{0,2}$	$-0,40^{0,1}$	$0,44^{0,1}$			
E5_R1	$2,61^{0,5}$	$-0,70^{0,2}$	$0,26^{0,1}$					
E6_R1	$2,55^{0,4}$	$-0,72^{0,2}$	$-0,08^{0,1}$	$0,52^{0,1}$	$0,86^{0,1}$	$1,23^{0,1}$	$1,85^{0,2}$	

Table 9.1: Polytomous IRT model item parameter estimates for the first exam

E1_R1		E2_R1		E3_R1		E4_R1		E5_R1		E6_R1	
Ca.	%	Ca.	%	Ca.	%	Ca.	%	Ca.	%	Ca.	%
0	0,20	0	0,11	0	0,07	0	0,10	0	0,25	0	0,25
1	0,13	1	0,09	1	0,06	1	0,05	1	0,31	1	0,19
2	0,27	2	0,12	2	0,05	2	0,18	2	0,44	2	0,21
3	0,40	3	0,13	3	0,11	3	0,28			3	0,11
		4	0,23	4	0,09	4	0,39			4	0,10
		5	0,22	5	0,25					5	0,09
		6	0,10	6	0,22					6	0,05
				7	0,15						

Table 9.2: Percent of students answered questions in the first exam w.r.t score categories (Ca. abbreviation of category)

Each exercise is made up of several dependent questions. The structure of both exams is almost the same. They are combinatorics, probability, frequency distribution, calculation of mean, standard deviation, covariance, density and distribution function. See Table 2.1 and 2.2 for the content of exams.

The software IRTPRO 2.1 for Windows (Li Cai, David Thissen & Stephen du Toit) was used to facilitate an analysis of polytomous IRT. It provides item parameters estimation, their standard errors, a measure of goodness of fit as well as testlet information functions.

The testlet (exercise) is considered as a unit and scored polytomously in order to make the local dependency within testlet disappear. We assume that all of the information about proficiency from each exercise is expressed by the summed score of correct items of that exercise. Each exercise j (j= 1, ..., 6) in the first exam has m_j questions (m_j = 3, 6, 7, 4, 2, 6). Each answer got equal and more than fifty percent of maximal points achieves a value of 1, otherwise 0. The answer with the value 1 is considered as correct. Exercise 1 is a three item testlet, the examinee´s score for this testlet can range from 0 to 3. An examinee has answered all questions of exercise 1 wrongly obtaining score 0 or answered all questions rightly getting score 3. Thus the first exercise has four score categories $x_1 = 0, 1, 2, 3$. The score categories of all six exercises in the first exam are 4, 7, 8, 5, 3 and 7, respectively.

The trace lines for six exercises of the first exam were shown in Figure 9.1. The trace line can be viewed as the regression of item score on the underlying variable θ. Each curve depicts each response category. The figure displayed which categories are less likely to be chosen. The more response categories each

Exercise	a^{SE}	b_1	b_2	b_3	b_4	b_5	b_6	b_7
E1_R2	$1,5^{0,3}$	$-2,54^{0,4}$	$-1,04^{0,2}$					
E2_R2	$0,99^{0,2}$	$-2,42^{0,4}$	$-0,1^{0,2}$					
E3_R2	$1,52^{0,2}$	$-3,5^{0,5}$	$-2,74^{0,4}$	$-1,97^{0,3}$	$-1,48^{0,2}$	$-0,63^{0,2}$	$0,55^{0,1}$	$1,78^{0,3}$
E4_R2	$2,26^{0,4}$	$-1,73^{0,2}$	$-0,94^{0,1}$	$-0,18^{0,1}$	$0,67^{0,1}$			
E5_R2	$1,39^{0,3}$	$-0,94^{0,2}$	$0,26^{0,1}$					
E6_R2	$2,04^{0,4}$	$-1,05^{0,2}$	$-0,66^{0,1}$	$-0,26^{0,1}$	$0,14^{0,1}$	$0,68^{0,1}$	$2,05^{0,2}$	

Table 9.3: Polytomous IRT item parameter estimates for the second exam

E1_R2		E2_R2		E3_R2		E4_R2		E5_R2		E6_R2	
Ca.	%	Ca.	%	Ca.	%	Ca.	%	Ca.	%	Ca.	%
0	0,05	0	0,11	0	0,02	0	0,09	0	0,26	0	0,20
1	0,18	1	0,36	1	0,02	1	0,12	1	0,29	1	0,09
2	0,77	2	0,53	2	0,05	2	0,22	2	0,45	2	0,11
				3	0,06	3	0,27			3	0,13
				4	0,16	4	0,30			4	0,16
				5	0,32					5	0,25
				6	0,25					6	0,06
				7	0,12						

Table 9.4: Percent of students answered questions in the second exam w.r.t score categories (Ca. abbreviation of category)

exercise has the more ambiguous the figure will be.

Taking a look at exercise 1, the probabilities of each of four response categories were presented in the upper left figure. The curve of the score 0 is a decreasing function of θ and it crosses the 0,5-probability line at $b_{i1} = -1,31$ (Figure 9.1 and Table 9.1). The curve of the highest score category $m_i = 3$ crosses the 0,5-probability line at $b_{i3} = 0,42$. The other curves do not have an evident relationship to the item parameters. The values b of the other exercises have the same characteristics. The exercises 2 and 6 seem to be the two "hardest" exercises with highest $b_{i6} = 1,74$ and 1,85. See section 3.5.1 for more details.

With increasing discrimination parameters a the curve will be steeper for the extreme score categories ($k = 0, m_i$). Meanwhile, the curves for the categories between the extremes ($k = 1, ..., m_{i-1}$) become more peaked. The values a of the last three exercises in the first exam are larger than the first three exercises. Hence, we could see the probabilities curves of exercises 4, 5, 6 more peaked than those of the others. The dash line in each figure is the information curve of each exercise (Figure 9.1). More about the information curve will be discussed in the next part. The trace lines for six exercises of the second exam were not shown here. Table 9.3 depicts the polytomous IRT parameter estimates of the second exam.

Generally, as parameter a increases, the probability of getting a specific score changes more quickly with a change in θ. The values b of the highest score categories of exercise 1, 2 are negative where the corresponding curves reach the 0,5-probability line. Exercises 1, 2 seem to be "easy". Exercises 3 and

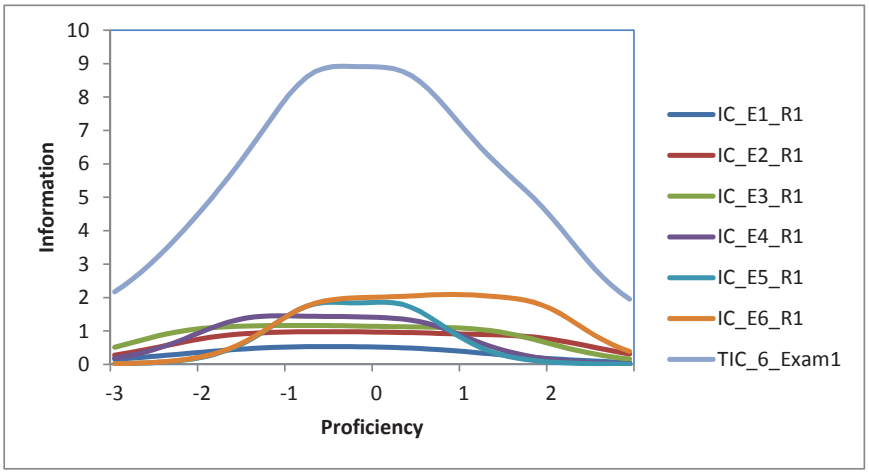


Figure 9.2: TIC and ICs of six exercises in the first exam

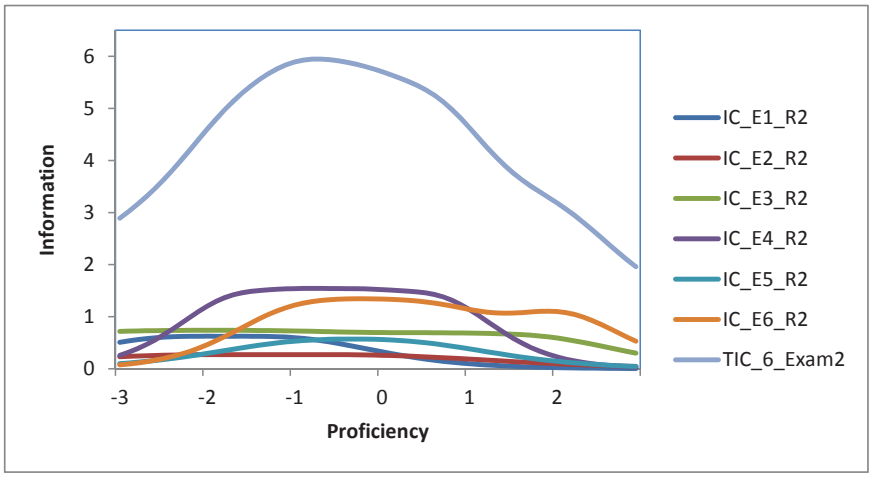


Figure 9.3: TIC and ICs of six exercises in the second exam

6 have the highest $b_{i7} = 1,78$ and $b_{i6} = 2,05$ which mean they are more difficult than the other ones. Table 9.2 and 9.4 showed the percent of students answering questions in the first and second exams with respect to the score categories. The score categories are the number of correct answers.

9.2 Information function

The information function (IF) indicates the precision of measurement for persons at different levels of proficiency. The shape of IF is dependent on the item

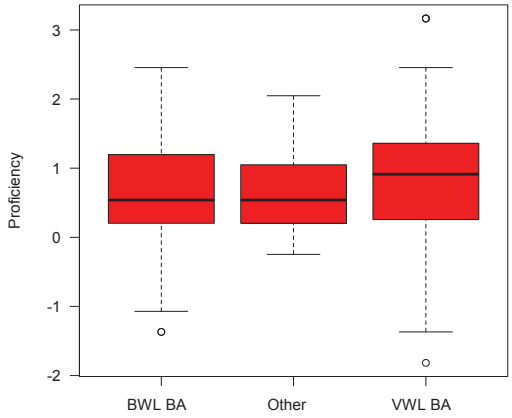


Figure 9.4: Boxplot of proficiency values of three groups in the first exam

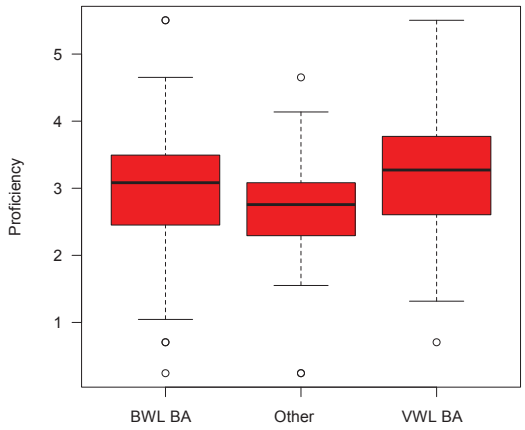
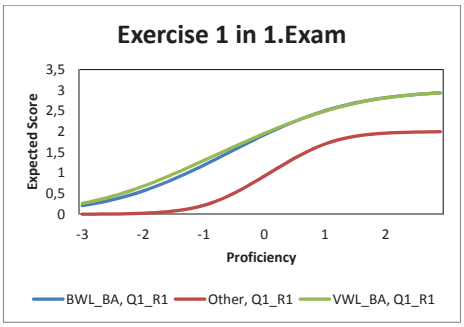


Figure 9.5: Boxplot of proficiency values of three groups in the second exam

parameters. The higher the item´s discrimination, the more peaked the IF will be. See section 3.5.4 for more details.

Figure 9.2 and 9.3 indicated information curves (ICs) for each of the testlets, as well as the total information curves (TICs) of six testlets for the first and second exams. Here we can see obviously the relationship between the information from the various exercises and the proficiency levels. The ICs of exercises 2, 3 and 4 are almost similar. Interestingly, exercise 6 of the first exam provides more information for examinees at the highest proficiency level. The curve spreads in the most area of high proficiency. Exercise 1 provides at least information, whereas exercise 5 yields peak information for examinees in the middle range of proficiency.

Exercise 2 gives at least information for the second exam, whereas exercise 1 brings more information at the lowest proficiency level. The information of exercise 3 is distributed uniformly across the proficiency range. Exercise 4 as well as exercise 6 yield much more information in the middle of proficiency axis



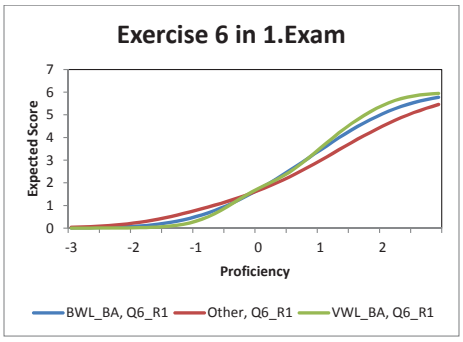
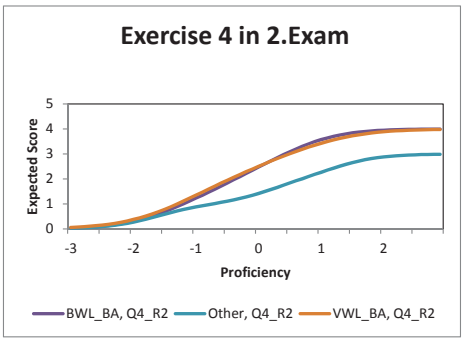


Figure 9.6: ES curves of exercise 1, 6 of three groups in the first exam



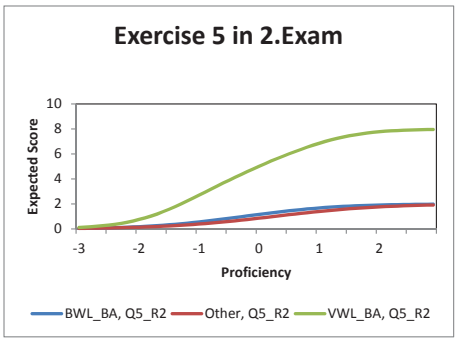


Figure 9.7: ES curves of exercise 4, 5 of three groups in the second exam

with $a_4 = 2,26$ and $a_6 = 2,04$ (Figure 9.3 and Table 9.3).

9.3 Expected Score

Figure 9.4 and 9.5 showed the boxplots of proficiency θ of all students in the first and second exams divided into three groups (BWL_BA, VWL_BA and Other). VWL_BA examinees have the highest proficiency values in both exams. Median of proficiency of BWL_BA students are larger than that of Other group.

Expected score (ES) functions of the polytomous IRT model can be considered analog to the item characteristic curves (ICCs) of dichotomously scored items. The ESs of all exercises were calculated with the formula 3.43. The expected scores of BWL_BA and VWL_BA of exercise 1 are much higher than those of the other group in the first exam. These values of exercise 1 in the second exam are nearly similar along the proficiency axis for three groups. In the first exam, there was very little differential item functioning (DIF) in expected scores within three groups in exercise 2, 3, 4. BWL_BA and VWL_BA examinees had higher expected scores on exercise 5 than the other group in the

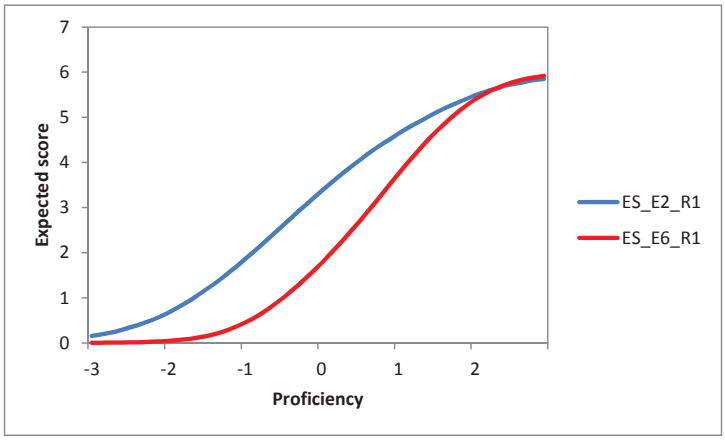


Figure 9.8: ES curves of exercise 2, 6 containing 6 questions in the first exam

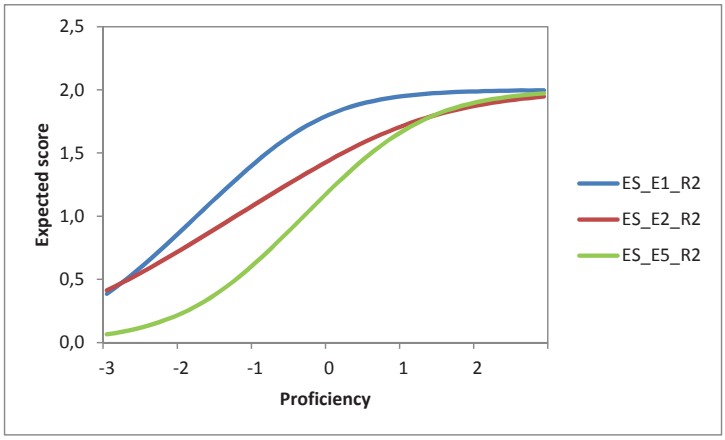


Figure 9.9: ES curves of exercise 1, 2 and 5 containing 2 questions in the second exam

first exam. VWL_BA students are obviously greater than the two other groups in exercise 5 in the second round at all levels of proficiency. See the ES figures of all exercises in both exams in Appendix 11.9.

Like the ICCs of dichotomously scored items, the ES curves of the polytomous IRT model indicate easy items if they shift more to the left and hard items if they shift more to the right.

Figure 9.8 presented the ESs of exercise 2, 6 composing of 6 questions in the first exam. The maximal value of ES is 6 (6 scores for 6 correct answers). The curve of exercise 6 shifts more to the right compared to that of exercise 2. It means that exercise 6 is more difficult than exercise 2. Figure 9.9 showed the ESs of exercise 1, 2 and 5 containing 2 questions in the second exam. The

No.	Exercise	χ^2	d.f.	Probability
1	E1_R1	45,95	37	0,15
2	E2_R1	76,08	57	0,05
3	E3_R1	89,86	58	0,0046
4	E4_R1	33,59	36	0,58
5	E5_R1	26,14	22	0,25
6	E6_R1	57,48	47	0,14

Table 9.5: S-χ^2 item level diagnostic statistics for the first exam

No.	Exercise	χ^2	d.f.	Probability
1	E1_R2	8,81	17	0,95
2	E2_R2	26,67	23	0,27
3	E3_R2	59,06	41	0,03
4	E4_R2	39,56	31	0,14
5	E5_R2	24,39	24	0,44
6	E6_R2	39,31	37	0,37

Table 9.6: S-χ^2 item level diagnostic statistics for the second exam

maximal score is 2. Exercise 5 seems to be the hardest one. The curve will move to the right with bigger b-parameters. A person with a particular proficiency level achieved more scores for exercise 1 than for exercise 5.

9.4 Goodness of fit tests

The χ^2 statistic tests the null hypothesis that an item (exercise) fits the polytomous IRT model. A nonsignificant chi-square value means that the null hypothesis can not be rejected. The null hypothesis for exercise 2, 3 in the first round was rejected at $\alpha = 5\%$. The rejection of H_0 for the other exercises failed which does not imply that the model is suitable for these exercises.

Only exercise 3 in the second exam is not consistent with the model at $\alpha = 5\%$. There is no sufficient evidence to reject the H_0 of the other exercises. Orlando and Thissen (2000, 2003) verified that all chi-square approaches are questionable. More about the fit analysis will be presented in section 10.

The marginal reliability of the GR model for the first and second exams are 0,86 and 0,80, respectively. They were computed with the formula 3.44.

Chapter 10

Comparison of 1PL, 2PL and polytomous IRT models

10.1 Comparison between 1PL and 2PL models

In this section, the obtained results of various IRT models will be compared to find out which model fits the data best. Table 10.1 depicts the values of difficulty parameters of 1PL and 2PL models in the first exam. The differences of b between two models are relative small. Almost values have the same sign. Negative values of location parameters mean “easy“ questions. Positive ones mean “hard“ questions. The larger the b values the more difficult the questions are. In short, the b estimates of 1PL model are little bit higher than those of 2PL model.

Table 10.2 presented the values of difficulty parameters of 1PL and 2PL models in the second exam. Several values of 1PL and 2PL models have different sign. Same to the first exam, b of 1PL model are also larger than b of 2PL model. Which model estimates parameters more accurate? The answer will be shown in the following.

The evaluation of fit in IRT modeling has been challenging (Embretson and Reise 2000). Many chi-square tests are computed for item or model fit. Orlando and Thissen (2000, 2003) indicated that all chi-square approaches are problematic. It is often not clear what the appropriate degree of freedom should be.

We would present information criteria and goodness-of-fit of 1PL and 2PL models. Table 10.3 showed Akaike (AIC), Bayesian Information Criterion (BIC), Deviance and McFadden´s R^2. See section 3.4.5 for more details.

Deviance and McFadden´s R^2 calculated only for 1PL model are used to measure and compare the fit of two models. The p-value of the deviance test in the first exam is 0,715, which will make a decision in favor of H_0. That means the first exam fits the 1PL model according to the value of deviance. Deviance D_0 of the second exam would reject the 1PL H_0 hypothesis with $p = 0,01$. R^2_{MF} in 1PL model of the first and second exam are 0,413 and 0,491, respectively. It is explained as proportional reduction in the deviance statistic (Menard 2000).

2PL model will become 1PL model when we impose the contraint that all items have the same discrimination parameters. Therefore, 1PL model is nested with 2PL model. Three statistics can be used to compare the nested models.

No.	Question	Theory/Practice	Field	1PL	2PL
1	Q41_R1	Practice	Bivariate	-2,47	-1,24
2	Q32_R1	Practice	Univariate	-1,63	-1,14
3	Q44_R1	Practice	Bivariate	-1,63	-0,83
4	Q31_R1	Practice	Univariate	-1,35	-1,26
5	Q33_R1	Practice	Univariate	-1,11	-0,87
6	Q22_R1	Theory	Probability	-1,06	-1,19
7	Q21_R1	Theory	Probability	-0,81	-0,8
8	Q12_R1	Theory	Combinatorics	-0,81	-1,02
9	Q11_R1	Theory	Combinatorics	-0,81	-0,99
10	Q36_R1	Practice	Univariate	-0,62	-0,52
11	Q52_R1	Practice	Bivariate	-0,58	-0,36
12	Q23_R1	Theory	Probability	-0,23	-0,46
13	Q61_R1	Theory	Univariate	-0,16	-0,29
14	Q43_R1	Practice	Bivariate	-0,07	-0,23
15	Q24_R1	Theory	Probability	-0,07	-0,38
16	Q37_R1	Practice	Univariate	0,02	-0,23
17	Q51_R1	Practice	Bivariate	0,33	0
18	Q35_R1	Practice	Univariate	0,33	-0,01
19	Q42_R1	Practice	Bivariate	0,36	0,01
20	Q25_R1	Theory	Probability	0,63	0,2
21	Q13_R1	Theory	Combinatorics	0,63	0,2
22	Q62_R1	Theory	Univariate	0,78	0,28
23	Q34_R1	Practice	Univariate	1,12	0,68
24	Q63_R1	Theory	Univariate	1,18	0,52
25	Q64_R1	Theory	Univariate	1,68	0,67
26	Q66_R1	Theory	Distribution	2,02	1,22
27	Q26_R1	Theory	Probability	2,06	1,39
28	Q65_R1	Theory	Univariate	2,23	0,95

Table 10.1: Comparison of difficulty parameters b of 1PL and 2PL models in the first exam

The first one based on the likelihood ratio test statistic is computed as follows

$$\Delta G^2 = -2ln(L_R) - (-2ln(L_F)) = G_R^2 - G_F^2$$

where L_R is the maximum of the likelihood for the reduced model and L_F is the maximum of the likelihood for the full model. The degrees of freedom of the test is the difference in the number of parameters between the full model and the reduced model.

As can be seen, the value of -2lnL is smaller for the more complex model. Taking the values -2lnL of 1PL and 2PL models from Table 10.3 we will obtain ΔG^2.

$$\Delta G^2 = -2ln(L_R) - (-2ln(L_F)) = 5000 - 4920,3 = 79,7$$

with $df = 27$. The second exam has the value of $\Delta G^2 = 166,6$ and $df = 22$. The 2PL model provided an improvement in fit over the 1PL model based on the significant p-value in two exams.

No.	Question	Theory/Practice	Field	1PL	2PL
1	Q31_R2	Practice	Univariate	-2,85	0,4
2	Q32_R2	Practice	Univariate	-2,69	-2,01
3	Q12_R2	Theory	Probability	-2,01	-1,91
4	Q41_R2	Practice	Bivariate	-2,01	-1,51
5	Q11_R2	Theory	Combinatorics	-1,09	-1,63
6	Q37_R2	Practice	Univariate	-0,53	-1,1
7	Q61_R2	Theory	Univariate	-0,49	-0,7
8	Q22_R2	Theory	Probability	-0,45	-1,23
9	Q36_R2	Practice	Univariate	-0,29	-0,95
10	Q21_R2	Theory	Probability	-0,03	-1,33
11	Q62_R2	Theory	Univariate	0,14	-0,33
12	Q52_R2	Theory	Bivariate	0,20	-0,52
13	Q42_R2	Practice	Bivariate	0,24	-0,5
14	Q43_R2	Practice	Bivariate	0,34	-0,39
15	Q33_R2	Practice	Univariate	0,56	-1,05
16	Q51_R2	Theory	Bivariate	0,71	-0,17
17	Q63_R2	Theory	Univariate	0,83	-0,01
18	Q34_R2	Practice	Univariate	1,01	0,11
19	Q65_R2	Theory	Univariate	1,02	0,07
20	Q35_R2	Practice	Univariate	1,20	0,36
21	Q44_R2	Practice	Bivariate	1,32	0,28
22	Q64_R2	Theory	Univariate	1,35	0,24
23	Q66_R2	Theory	Distribution	3,43	3,95

Table 10.2: Comparison of difficulty parameters b of 1PL and 2PL models in the second exam

The second test statistic is shown in the following equation.

$$R^2_\Delta = \frac{(G^2_R - G^2_F)}{G^2_R}$$

$$R^2_\Delta = \frac{(G^2_R - G^2_F)}{G^2_R} = \frac{5000 - 4920,3}{5000} = 0,016$$

$R^2_\Delta = 0,043$ is the value of the second round which pointed out that the 2PL model resulted in a 1,6% and 4,3%, respectively improvement in fit over the 1PL model.

The third approach used primarily for model comparisons is Akaike information criterion (AIC) or Bayesian information criterion (BIC). For competing models, the model which minimizes an information criteria is normally selected.

$$AIC = -2lnL + 2n$$

$$BIC = -2lnL + ln(m)n$$

where n is the number of estimated parameters, m is the number of persons. In the second exam, the AIC and BIC are smaller for the 2PL model (Table

Index	1.Exam		2.Exam	
	1PL	2PL	1PL	2PL
-2lnL	5000,0	4920,3	3866,6	3700,0
Akaike Information Criterion (AIC)	5056,0	5032,3	3912,6	3792,0
Bayesian Information Criterion (BIC)	5144,8	5209,8	3984,9	3936,5
Deviance	705,86		533,50	
McFadden R^2	0,41		0,49	

Table 10.3: Goodness-of-fit tests in 1PL, 2PL models for the first and second exams

10.3). There is a conflict in determination of choosing the model in the first exam since BIC is larger for the 2PL model. In summary, we could state that the 2PL model indicated the better fitting.

10.2 Comparison between 2PL and polytomous IRT models

With the assumption of local independence within items in the 2PL IRT model, the TIC can be created from the item information of all the items. At each level of the underlying trait, the IF is approximately equal to the expected value of the inverse of the squared standard errors of the θ estimates (Lord, 1980).

Figure 10.1 and 10.2 showed the results of TIC and SE of 2PL and polytomous IRT models in both exams. The 2PL model yielded much more information than the polytomous model over most values of proficiency in the first exam. The blue line is for the 2PL IRT model. The red line is for the polytomous IRT model. Up to nearly θ=1 the 2PL model provided more information in the second round. That implied the polytomous IRT model gave a more accurate estimation for high-level proficiency.

In polytomous IRT model, only the total score are used to describe the attribute of the data. That is the reason of losing information. The 2PL model gives more information due to the conditional independence assumption. Our data are not mutually independent. The response to the following item depends on the response to the previous item in some exercises, which makes the following item provide less information about proficiency than a completely independent item. It wil be not accurate to fit the data with the 2PL model. Hence, we can state here that the polytomous IRT model despite providing less information is a better approach for the testlet-based data.

In summary, IFs can be used to design an instrument with some particular characteristics. Tester can select adequate items based solely on the item parameter estimates so that they can get the maximum information as they want.

Until now, only the estimation accuracy at different points along the continuum has been indicated. We also need a single bounded value representing the quality of estimation for the entire continuum, such as empirical reliability,... The empirical reliability based on the ratio of the variance of the expected a posteriori (EAP) $\hat{\theta}s$ to the sum of the variance of the $\hat{\theta}s$ and error variance (Zimowski et al., 2003) ranges from 0 to 1. With the value nearly or equal to 1,

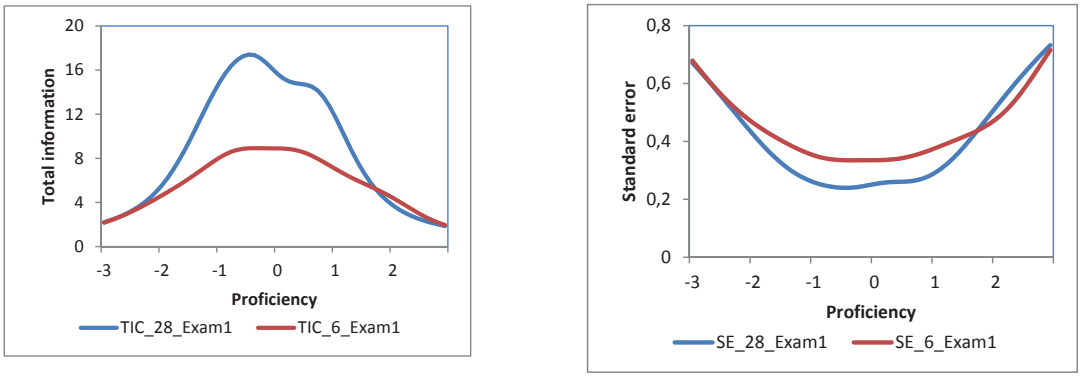


Figure 10.1: TIC & SE for 2PL IRT model and polytomous IRT model in the first exam

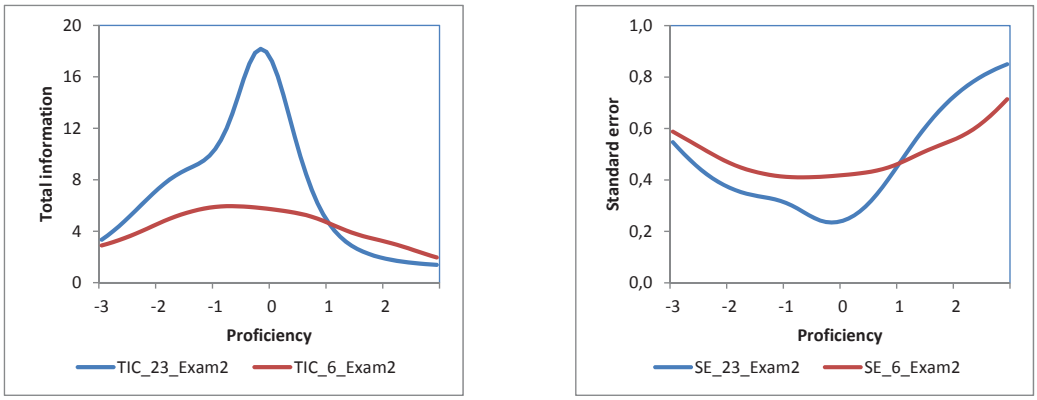


Figure 10.2: TIC & SE for 2PL IRT model and polytomous IRT model in the second exam

the model is considered as a good one. This index of the first and second exams in the 2PL IRT model are 0,91 and 0,86, respectively. Those values lie in the acceptable interval reflecting small error variability. The marginal reliability of the GR model for the first and second exams are 0,86 and 0,80 computed with the formula 3.44.

Chapter 11

Conclusion

In this chapter, we will provide an overview of the obtained results and several drawn conclusions. First of all, the reliability analysis determined by Cronbach´s α indicated that the two statistics exams measured the single latent contruct in a great manner (Table 5.1). The values of Cronbach´s α have not increased when almost any of the items in both data sets was dropped.

Explanatory factor analysis based on tetrachoric correlation was performed for two exams. The interrelationships with the items have been explored and explained by a small number of latent variables. Table 6.2 and 6.6 contain the values of factor loadings in the two-factor model. In both exams, about 45-46% of total variance was explained through the first factor. Nine questions of exercises 1 and 2 are loaded on the first factor, the combinatorics-probability factor. The exercises 3, 4, 5 and 6 are loaded on the second factor, the univariate-bivariate factor. Almost questions of exercises 1, 4, 5 and 6 in the second exam are loaded on the first factor, the theoretical factor. The second factor has strong loadings on nearly all questions of exercises 2 and 3, the practical factor. In summary, we can state that the factor one is about proficiency in theory, the factor two is about proficiency in application.

In confirmatory factor analysis the Chi-squared test and goodness of fit indices such as RMSEA, TLI and CFI were computed to evaluate to what extent a particular factor model fits the empirical data. These values have again confirmed that the two-factor model explained the observed data of the second exam consistently. However, for the first exam only the five-factor model seems to be a good model based on the RMSEA, TLI and CFI (Table 7.1 and 7.2).

Item response theory is a useful tool for both test theory and test developement. Nevertheless, the assumption of local independence within items makes the 1PL and 2PL models become limited in several types of test. The 1PL item response theory model characterizes each item in term of a single parameter, difficulty parameter. The difficulty parameters of 28 items in the first exam spread over the range of ability axis which means examinees of all level abilities measured well by items with all levels of difficulty. There was a lack of several "harder" items for high-proficiency examinees in the second round. That is possibly the intention of the testers to let more students pass the exam (Figure 8.1 and 8.2).

The 2PL item response theory model allows for different slopes. As a rule, the items possessing extremely high or very low discrimimation parameters vi-

olate the local independency assumption of the dichotomous model and are frequently not loaded on one single factor in explanatory factor analysis. Most items in the first exam with reasonable values of slope parameters have high enough the separation power. Two questions of exercise 6 in the second exam have the slopes exceeding 4 which implied the violation of the local independence assumption (Table 8.1 and 8.2).

The polytomous item response theory was applied to testlet-based data using IRTPRO 2.1. The items in the statistics exams are not mutually independent. The response to the following item depends on the response to the previous item in some exercises. In graded response model, each exercise is scored polytomously in order to make the local dependency within the exercise disappear. The obtained outcomes showed that exercises 2 and 6 in the first exam are the two "hard" exercises. Exercises 1 and 2 in the second exam seem to be "easy". Exercises 3 and 6 have the higher difficulty parameters which mean they are more difficult than the other ones (Table 9.1 and 9.3).

Exercise 1 in the first exam gives at least information, but on the other hand exercise 5 yields peak information for examinees in the middle range of proficiency. The information curves of exercises 2, 3 and 4 are not much different. Exercise 6 provides more information for examinees at the rather high-proficiency level. For the second exam exercise 2 gives at least information, whereas exercise 1 brings more information at the low-proficiency level. The information of exercise 3 is distributed uniformly across the proficiency range. Exercise 4 as well as exercise 6 yield much more information in the middle of proficiency axis (Figure 9.2 and 9.3).

The result of deviance test was not sufficient to reject the null hypothesis that the first statistics exam fits the 1PL model. With $p = 0,01$ the H_0 for the second one would be rejected. According to Akaike and Bayesian information criterion, we could conclude that the 2PL model indicated a better fitting in two exams.

The comparison of the 2PL and polytomous models was carried out based on the total information and standard error. The 2PL model provides much more information than the polytomous model over most values of the proficiency range in the first exam. For the second exam, the polytomous model yields a more accurate parameters estimation for high-level proficiency (Figure 10.1 and 10.2).

We also would like to implement the analysis of the statistics data using testlet models with the computer programm Scoright 3.0 developed by H. Wainer, X. Wang and E. Bradlow. Unfortunately, the program has not functioned well with our real data. Maybe the small sample size of the first and second exams is a factor influencing to the process of the parameters estimation.

Appendix

No.	Question	Theory/Practice	Field	F1	F2	F3	F4
1	Q11_R2	Theory	Combinatorics	0,41			
2	Q12_R2	Theory	Probability		0,91		
3	Q21_R2	Theory	Probability		0,42		
4	Q22_R2	Theory	Probability			0,36	
5	Q31_R2	Practice	Univariate			0,61	
6	Q32_R2	Practice	Univariate			0,83	
7	Q33_R2	Practice	Univariate		0,49		
8	Q34_R2	Practice	Univariate			0,53	
9	Q35_R2	Practice	Univariate				0,57
10	Q36_R2	Practice	Univariate			0,89	
11	Q37_R2	Practice	Univariate			0,92	
12	Q41_R2	Practice	Bivariate				0,60
13	Q42_R2	Practice	Bivariate				0,59
14	Q43_R2	Practice	Bivariate				0,63
15	Q44_R2	Practice	Bivariate				0,45
16	Q51_R2	Theory	Bivariate		0,47		
17	Q52_R2	Theory	Bivariate		0,41		
18	Q61_R2	Theory	Univariate	0,89			
19	Q62_R2	Theory	Univariate	0,83			
20	Q63_R2	Theory	Univariate	0,72			
21	Q64_R2	Theory	Univariate	0,81			
22	Q65_R2	Theory	Univariate	0,88			
23	Q66_R2	Theory	Distribution	0,35			

Table 11.1: Four-factor model of the second exam

No.	Question	Theory/Practice	Field	F1	F2	F3	F4	F5
1	Q11_R2	Theory	Combinatorics		0,43			
2	Q12_R2	Theory	Probability		0,91			
3	Q21_R2	Theory	Probability			0,65		
4	Q22_R2	Theory	Probability				0,40	
5	Q31_R2	Practice	Univariate				0,74	
6	Q32_R2	Practice	Univariate				0,98	
7	Q33_R2	Practice	Univariate		0,54			
8	Q34_R2	Practice	Univariate				0,64	
9	Q35_R2	Practice	Univariate				0,51	
10	Q36_R2	Practice	Univariate				0,73	
11	Q37_R2	Practice	Univariate				0,70	
12	Q41_R2	Practice	Bivariate				0,70	
13	Q42_R2	Practice	Bivariate	0,51				
14	Q43_R2	Practice	Bivariate	0,78				
15	Q44_R2	Practice	Bivariate	0,49				
16	Q51_R2	Theory	Bivariate		0,48			
17	Q52_R2	Theory	Bivariate		0,49			
18	Q61_R2	Theory	Univariate					0,88
19	Q62_R2	Theory	Univariate					0,82
20	Q63_R2	Theory	Univariate					0,72
21	Q64_R2	Theory	Univariate					0,81
22	Q65_R2	Theory	Univariate					0,89
23	Q66_R2	Theory	Distribution					0,35

Table 11.2: Five-factor model of the second exam

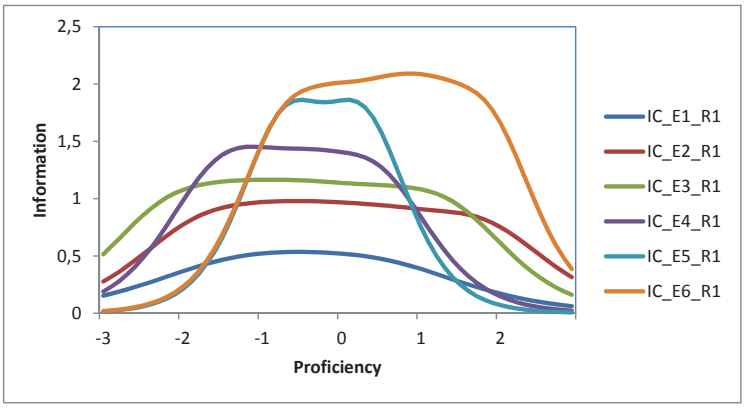


Figure 11.1: IC of six exercises in the first exam

Item	Question	χ^2	d.f.	Probability
1	Q11_R1	19,75	15	0,18
2	Q12_R1	15,8	17	0,54
3	Q13_R1	15,55	13	0,27
4	Q21_R1	16,24	16	0,44
5	Q22_R1	19,97	16	0,22
6	Q23_R1	20,16	16	0,21
7	Q24_R1	22,2	14	0,07
8	Q25_R1	15,65	14	0,34
9	Q26_R1	6,42	13	0,93
10	Q31_R1	26,02	14	0,03
11	Q32_R1	19,58	12	0,08
12	Q33_R1	14,69	14	0,4
13	Q34_R1	12,32	14	0,58
14	Q35_R1	20,11	16	0,21
15	Q36_R1	16,82	12	0,16
16	Q37_R1	17,21	15	0,31
17	Q41_R1	9,8	5	0,08
18	Q42_R1	9,21	14	0,82
19	Q43_R1	13,26	13	0,43
20	Q44_R1	12,33	7	0,09
1	Q51_R1	11,92	14	0,61
22	Q52_R1	8,26	8	0,41
23	Q61_R1	15,39	13	0,28
24	Q62_R1	18,63	12	0,1
25	Q63_R1	13,91	12	0,31
26	Q64_R1	16,87	9	0,05
27	Q65_R1	5,92	7	0,55
28	Q66_R1	18,41	14	0,19

Table 11.3: S-χ^2 item statistics of 2PL IRT model for the first exam

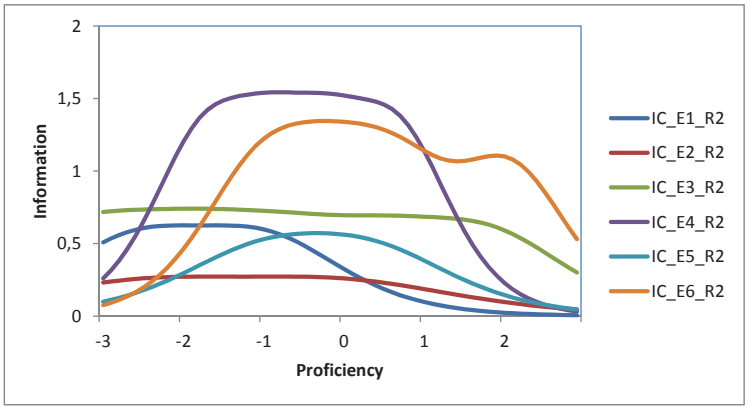


Figure 11.2: IC of six exercises in the second exam

Item	Question	χ^2	d.f.	Probability
1	Q11_R2	4,71	13	0,98
2	Q12_R2	7,41	8	0,49
3	Q21_R2	16,76	16	0,4
4	Q22_R2	11,51	14	0,65
5	Q31_R2	4	5	0,55
6	Q32_R2	3,57	4	0,47
7	Q33_R2	23,81	15	0,07
8	Q34_R2	18,73	14	0,17
9	Q35_R2	16,92	14	0,26
10	Q36_R2	18,63	12	0,1
11	Q37_R2	13,02	13	0,45
12	Q41_R2	6,47	7	0,49
13	Q42_R2	13,3	13	0,43
14	Q43_R2	19,73	12	0,07
15	Q44_R2	9,03	11	0,62
16	Q51_R2	14,18	13	0,36
17	Q52_R2	12,6	11	0,32
18	Q61_R2	9,87	8	0,28
19	Q62_R2	5,01	6	0,54
20	Q63_R2	6,61	7	0,47
21	Q64_R2	15,97	8	0,04
22	Q65_R2	7,68	8	0,47
23	Q66_R2	5,53	8	0,7

Table 11.4: S-χ^2 item statistics of 2PL IRT model for the second exam

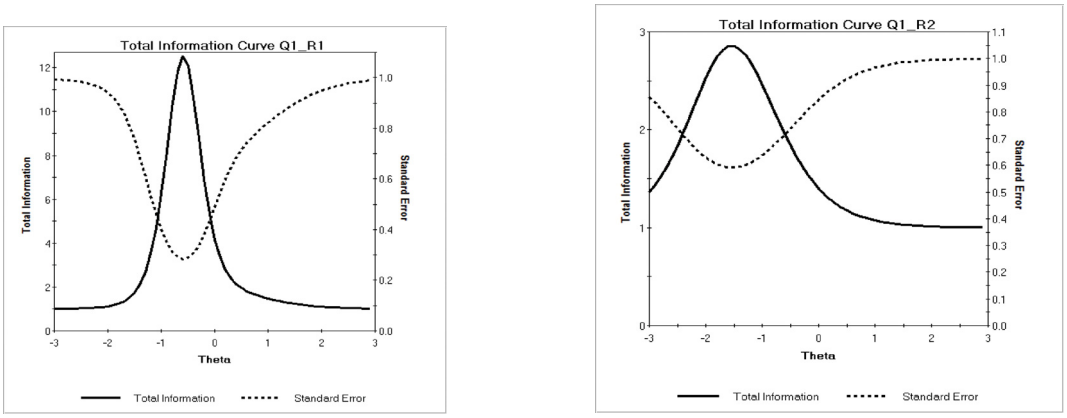
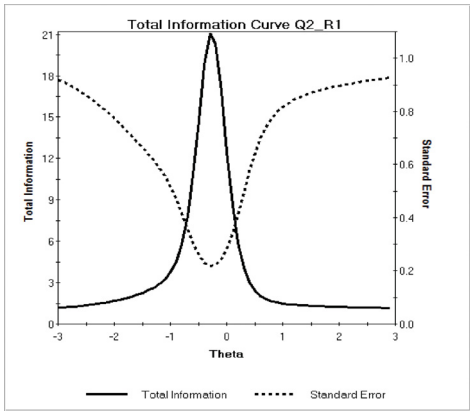


Figure 11.3: TIC and SE of 2PL model for all questions of exercise 1 in the first and second exams



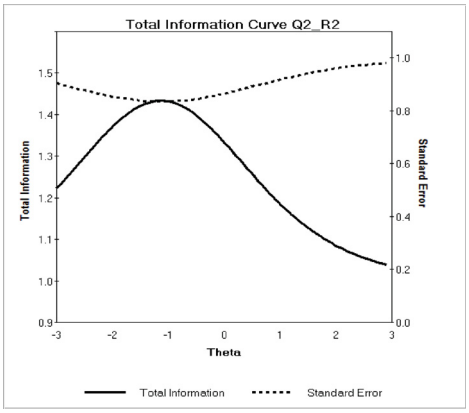
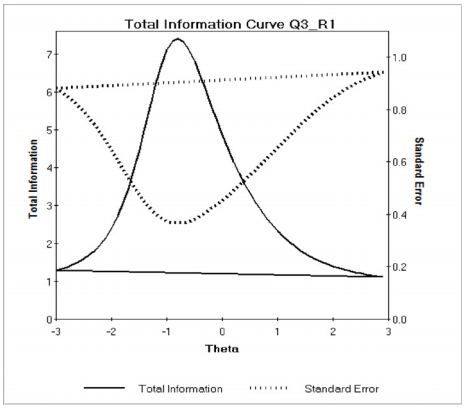


Figure 11.4: TIC and SE of 2PL model for all questions of exercise 2 in the first and second exams



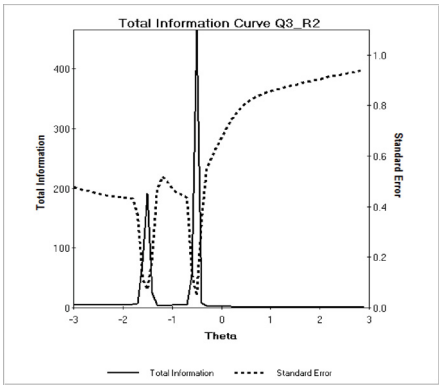


Figure 11.5: TIC and SE of 2PL model for all questions of exercise 3 in the first and second exams

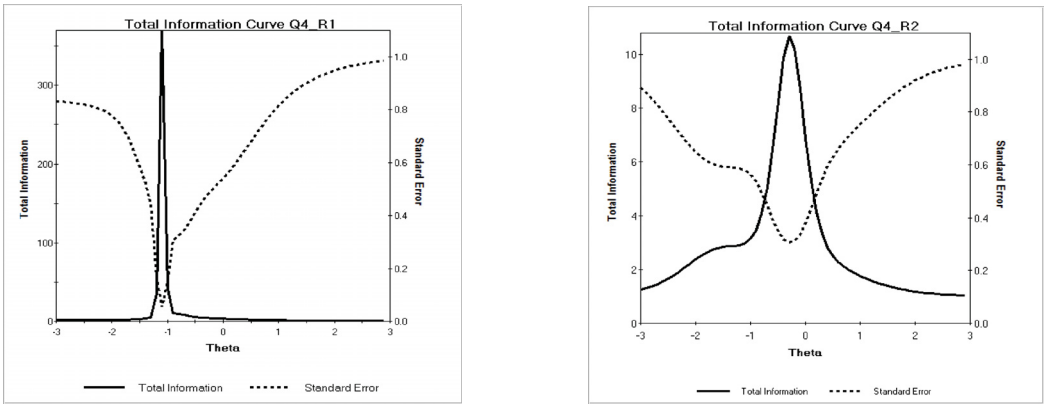


Figure 11.6: TIC and SE of 2PL model for all questions of exercise 4 in the first and second exams

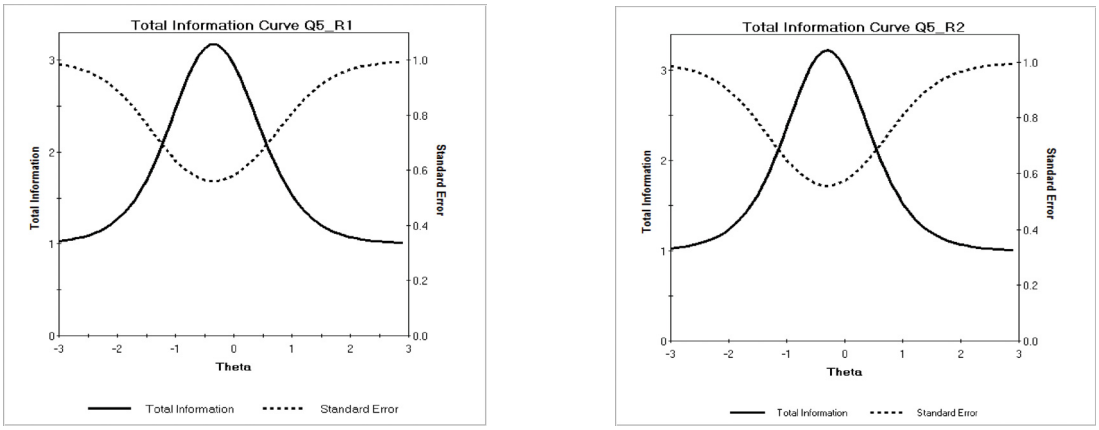


Figure 11.7: TIC and SE of 2PL model for all questions of exercise 5 in the first and second exams

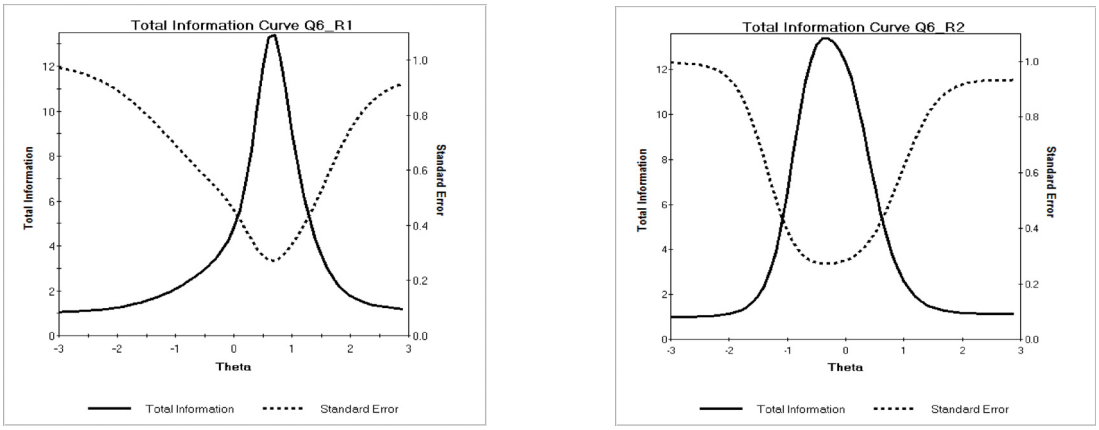


Figure 11.8: TIC and SE of 2PL model for all questions of exercise 6 in the first and second exams

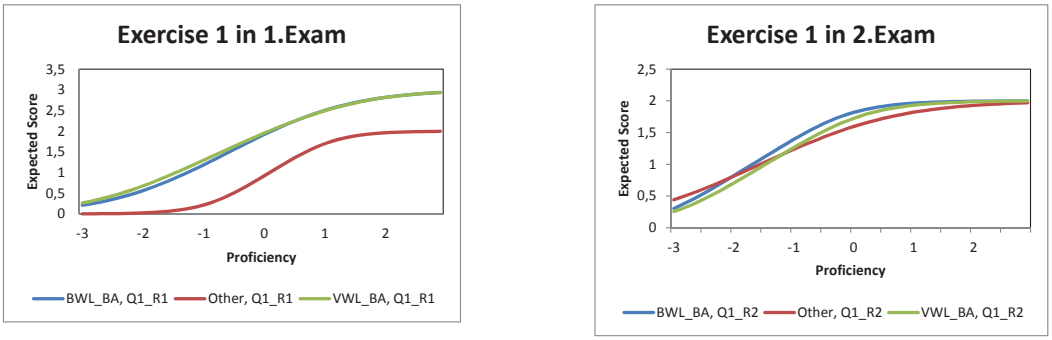
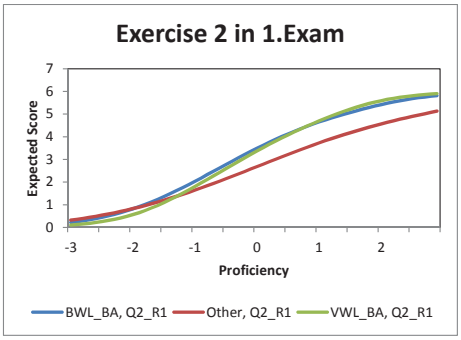


Figure 11.9: ES curves of exercise 1 of three groups in the first and second exams



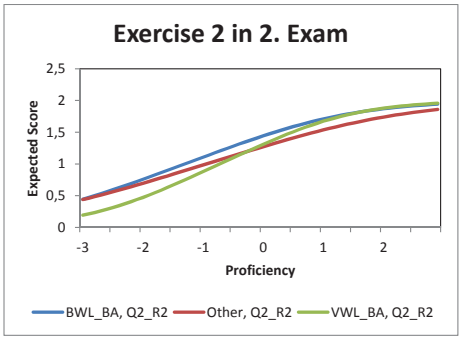
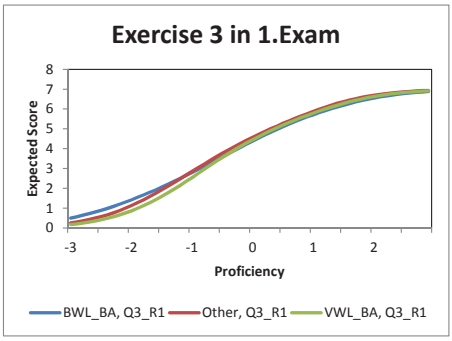


Figure 11.10: ES curves of exercise 2 of three groups in the first and second exams



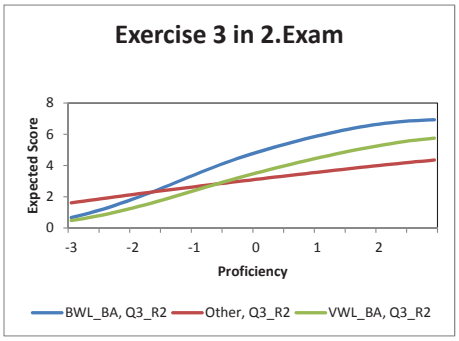
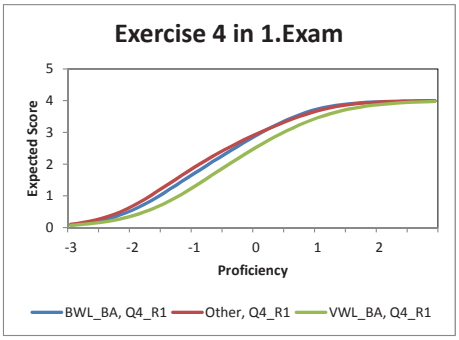


Figure 11.11: ES curves of exercise 3 of three groups in the first and second exams



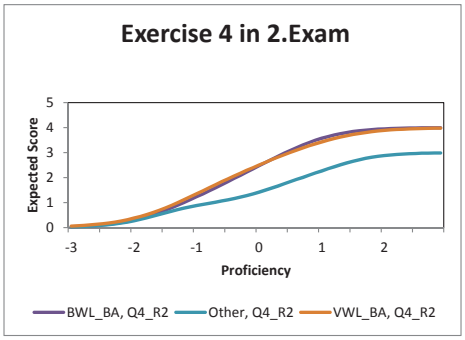


Figure 11.12: ES curves of exercise 4 of three groups in the first and second exams

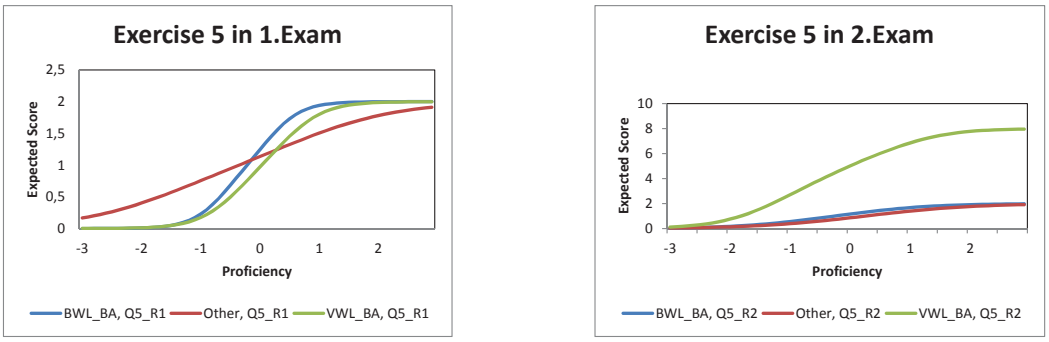


Figure 11.13: ES curves of exercise 5 of three groups in the first and second exams

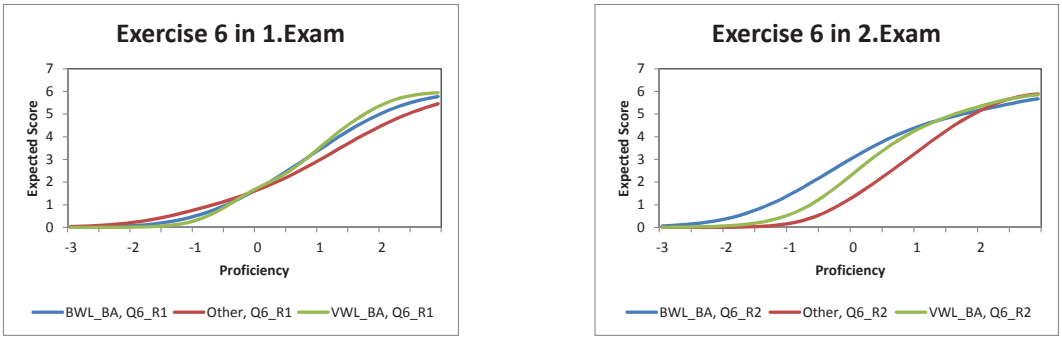


Figure 11.14: ES curves of exercise 6 of three groups in the first and second exams

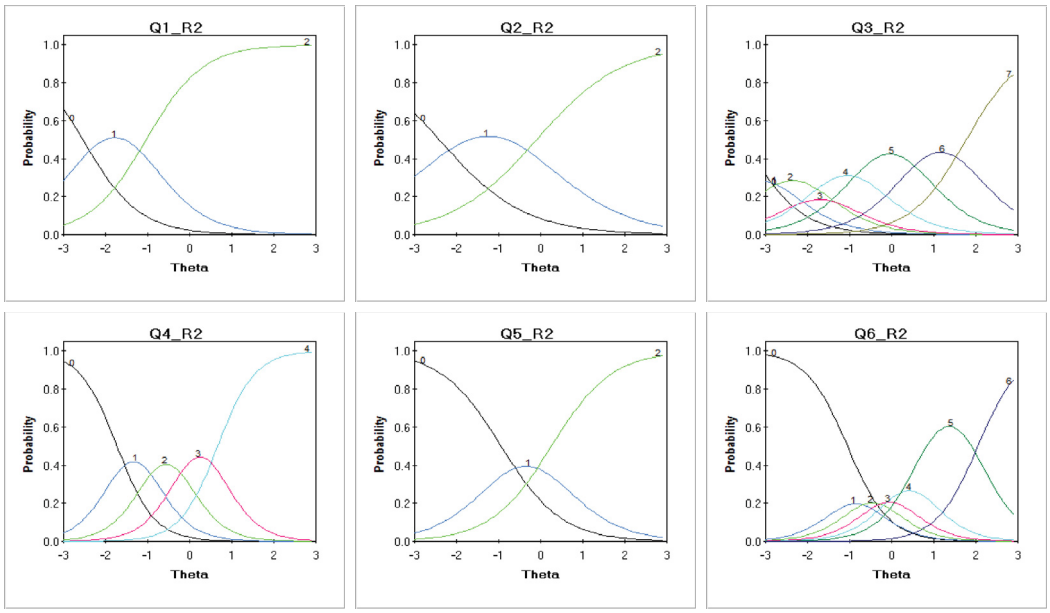


Figure 11.15: Trace lines of 6 exercises in the second exam

Bibliography

Ayala, R. J. (2009). *The theory and practice of item response theory*, The Guilford Press New York London.

Ayearst, L. E. and Bagby, M. (2011). *Evaluating the psychometric properties of psychological measures*, from Handbook of Assessment and Treatment Planning for Psychological Disorders edited by M. Antony and H. Barlow, Guilford Press.

Bartholomew, D., Steele, F., Moustaki, I. and Galbraith, J. I. (2000). *The analysis and interpretation of multivariate data for social scientists*, Chapman & Hall/CRC.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee´s ability, *Statistical theories of mental test scores* pp. 392–479.

Boeck, P. D. and Wilson, M. (2004). *Explanatory Item Response Models A Generalized Linear and Nonlinear Approach*, Spinger.

Embretson, S. E. and Reise, S. P. (2000). *Item response theory for psychologists*, Lawrence Erlbaum Associates, Publishers.

Fischer, G. H. and Molenaar, I. W. (1995). *Rasch Models Foundations, Recent Developments, and Applications*, Springer Verlag.

Härdle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*, Springer.

Johnson, M. S. (2007). Marginal Maximum Likelihood Estimation of Item Response Models in R, *Journal of Statistical Software* .

Kline, P. (2000). *The new psychometrics: science, psychology and measurement*, Routledge.

Klinke, S. and Wagner, C. (2008). Visualizing exploratory factor analysis models, *In Sonderforschungsbereich 649: Ökonomisches Risiko, Humboldt-Universität zu Berlin* .

Linden, W. J. and Hambleton, R. K. (1997). *Handbook of modern item response theory*, Spinger.

Mair, P. and Bentler, P. M. (2008). IRT Goodness-of-Fit Using Approaches from Logistic Regression, *Department of Statistics, UCLA, UC Los Angeles* .

Muthén, B. O. and Lehman, J. (1985). Multiple-group item response theory modeling: Applications to item bias analysis, *Journal of Educational Statistics* .

Muthén, B. O., Toit, S. and Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.

Reckase, M. D. (2009). *Multidimensional item response theory*, Spinger.

Rizopoulos, D. (2012). Latent Trait Models under IRT, *Package ltm in R* .

Szabó, G. (2008). *Applying item response theory in language test item bank building*, Lang.

Tang, K. L. (1996). Polytomous Item Response Theory Models and Their Applications in Large-scale Testing Programs.

Wainer, H., Bradlow, E. and Wang, X. (2005). User's guide for scoright 3.0: A computer program for scoring tests built of testlets including a module for covariate analysis.

Wainer, H., Bradlow, E. and Wang, X. (2007). *Testlet Response Theory and Its Applications*, Cambridge University Press.